# Low-Resolution RIS-Aided Multi-User MIMO Signaling

A. A. Nasir, H. D. Tuan, E. Dutkiewicz, H. V. Poor, and L. Hanzo

***Abstract*—A multi-antenna aided base station (BS) supporting several multi-antenna downlink users with the aid of a reconfigurable intelligent surface (RIS) of programmable reflecting elements (PREs) is considered. Low-resolution PREs constrained by a set of sparse discrete values are used for reasons of cost-efficiency. Our challenging objective is to jointly design the beamformers at the BS and the RIS's PREs for improving the throughput of all users by maximizing their geometric-mean, under a variety of different access schemes. This constitutes a computationally challenging problem of mixed continuous-discrete optimization, because each user's throughput is a complicated function of both the continuous-valued beamformer weights and of the discrete-valued PREs. We develop low-complexity algorithms, which iterate by directly evaluating low-complexity closed-form expressions. Our simulation results show the advantages of non-orthogonal multiple access-aided signaling, which allows the users to decode a part of the multi-user interference for enhancing their throughput.**

***Index Terms*—Reconfigurable intelligent surface, coordinated signaling (CoSig), non-orthogonal multiple access (NOMA), MIMO beamforming, trigonometric function optimization, geometric mean maximization, low-complex algorithms**

## I. Introduction

Reconfigurable intelligent surfaces (RISs) relying on large numbers of programmable reflecting elements (PREs), have recently been shown to beneficially improve the wireless coverage and security [1]–[3]. A hot topic in RIS-aided multi-user (MU) communication is the joint design of the downlink (DL) transmit beamformers at the base station (BS) and the RIS PREs for spectral efficiency in terms of the sum-throughput (ST) subject to transmit power constraints [4]–[7], the transmit power subject to the signal-to-interference-plus-noise-ratio (SINR) constraints [8], [9], or users' individual throughput [10], [11]. All these works rely on multiple input single output (MISO) schemes, since the users are equipped with single antennas. Furthermore, there is a paucity of literature on low-resolution RIS PREs due to their discrete-valued natures.

The explosive growth of internet-of-things (IoT) applications motivates the aggressive reuse of the resources, such as the bandwidth and time slots to serve as many users as possible. Hence efficient multi-user interference (MUI) management is critical for maintaining a high quality-of-service (QoS) in terms of information throughput. Non-orthogonal multiple access (NOMA) [12]–[14] may be viewed as a particular class of MU multiple input single output (MISO) transmit beamforming, which enables the clustered users to successively decode the whole messages destined to other users of the same cluster to mitigate intra-cluster interference for their message throughput enhancement. In this context, it is important to compare its spectral efficiency to that achieved by the conventional coordinated signaling scheme (CoSig), under which each user decodes its own message by treating the MUI as noise. The most popular NOMA scheme pairs users, so both of the two users of the same cluster decode the entire message destined for one of them. In terms of users' worst throughput, this NOMA scheme was shown to outperform CoSig, provided that the channel conditions of the paired users are sufficiently differentiated [15]. However, when the paired users have similar channel conditions, the former is outperformed by the latter [16], [17]. A new-NOMA (n-NOMA) scheme was conceived in [17], which assigns only a part of the message intended to one user, as the common message to be decoded by both of them. This solution has been shown to outperform both NOMA and CoSig under flexible users' channel conditions.

It has recently been shown that RISs are also capable of supporting NOMA [18], [19]. The performance analysis of RIS-aided MISO-NOMA networks has been studied in [20]–[22]. Resource allocation in RIS-aided NOMA communication, which relies on the joint design of the transmit beamformers at the base station (BS) and PREs at the RIS, has been studied in [23], [24] for single-input-single-output (SISO) cases, and in [25]–[28] for MISO cases. Their computation is based on semi-definite relaxation, which is computationally demanding due to its dependency on the solution of a high-dimensional convex problem at each iteration. Yet, the convergence of their algorithms is not guaranteed [29]. Their spectral efficiency comparison to conventional non-optimal orthogonal multiple access (OMA) (with no optimal time or bandwidth allocations) cannot be conclusive because these OMA schemes are attractive owing to their simple single-user detection, but

The work was supported in part by the Deanship of Research Oversight and Coordination (DROC) at KFUPM for funding under the Interdisciplinary Research Center for Communication Systems and Sensing through project No. INCS2203, in part by the Australian Research Council's Discovery Projects under Grant DP190102501, in part by the U.S. National Science Foundation under Grant CNS-2128448, in part by the Engineering and Physical Sciences Research Council projects EP/W016605/1 and EP/P003990/1 (COALESCE) as well as by the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028).

A. A. Nasir is with the Department of Electrical Engineering and Center for Communication Systems and Sensing at King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia (email: anasir@kfupm.edu.sa).

H. D. Tuan and E. Dutkiewicz are with the School of Electrical and Data Engineering, University of Technology Sydney, Broadway, NSW 2007, Australia (email:Tuan.Hoang@uts.edu.au, Eryk.Dutkiewicz@uts.edu.au).

H. V. Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA (email: poor@princeton.edu).

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

their spectral efficiency has not been optimized.[1] Furthermore, these previous researches cannot address quantized RIS PREs and cannot be applied to RIS-aided multiple-input multiple-output (MIMO)-NOMA systems.

Against the above background, this paper is the first work to consider RIS-aided MU MIMO networks relying on low-resolution RIS PREs. It develops a novel low-complexity design for DL transmit beamformers at the BS and low-resolution constrained PREs at the RIS to maximize the geometric-mean of the users' throughput (GM-throughput). Our recent paper [11] shows that the GM-throughput constitutes a more appropriate metric, since its optimization naturally leads to fair throughput distribution associated with low standard deviation among the users' throughput without enforcing nonconvex throughput constraints. It is thus in contrast to the ST, whose maximization results in unfair throughput distribution by allocating a major portion of the resources to a few users having favorable channel conditions, while leaving the users having low channel quality with near-zero throughput. In contrast to the minimum users' throughput, which is a non-smooth function, the GM-throughput is smooth and thus lends itself to developing convenient computationally tractable solutions. In contrast to our previous research [10], [11], this work considers (i) MU MIMO networks; (ii) RIS-aided communication under NOMA, n-NOMA, and CoSig; (iii) low-resolution constrained PREs for practical implementation.

The advantage of adopting these novel elements will also be shown through extensive simulations. Explicitly, our new contributions are as follows:

- The design under both convex sum-power constraint and nonconvex individual transmit-antenna (TA) power constraints is considered, while the RIS PREs are constrained by discrete values of their low resolution.
- The transmit beamformers at the BS and the PREs at the RIS are optimized for the RIS-aided MIMO network by iterating the linearly scalable closed-form expressions. As such, our results are novel even considered from mathematical programming.
- We assess the performance of the proposed algorithms through extensive simulations by varying the transmit power budget, the number of PREs, the resolution of the phase shifters in the PREs, and the number of receiver antennas. The performance of the proposed algorithms is also evaluated for another scenario, where a direct communication between BS and UEs is blocked by obstacles. Explicitly,
  - The simulation results show the superiority of n-NOMA and NOMA-based RIS implementation over CoSig. However, for a particular scenario, where direct communication between the BS and UEs is blocked, NOMA is outperformed by CoSig under per-TA power constraints based design.
  - Compared to the conventional sum-throughput maximization, the achievable sum-throughput of the proposed GM-throughput maximization algorithms is smaller, as expected. However, it is shown that the achievable sum-throughput of n-NOMA is only slightly compromised compared to the NOMA and CoSig schemes.

[1]To the best of our knowledge, only [30] provides a comparison between NOMA and optimal OMA.

TABLE I: Boldly and explicitly contrasting our contributions to the literature

| | **this work** | [4]–[9] | [10] | [11] | [25]–[28] |
|---|---|---|---|---|---|
| MU MIMO | ✓ | | | | |
| $b$-bit PREs optimization | ✓ | | | | |
| NOMA | ✓ | | | | ✓ |
| n-NOMA | ✓ | | | | |
| CoSig | ✓ | ✓ | ✓ | ✓ | |
| Semi-definite relaxation | | ✓ | | | ✓ |
| Computational tractability | ✓ | | | ✓ | |

In a nutshell, we boldly contrast our new contributions to the state-of-the-art in Table I.

The paper is organized as follows. Section II is devoted to the problem formulation of GM-throughput optimization to jointly design the transmit beamformers and RIS's PREs. Section III derives a low-complexity solution of the problem formulated. Section IV is dedicated to solving the same problem under individual TA power constraints instead of a total sum-power constraint. Section V provides our detailed numerical simulations for quantifying the performance of the proposed algorithms. Finally, Section VI concludes the paper.

*Notation.* Only the beamformer and PRE variables are printed in boldface; $I_n$ is the identity matrix of size $n \times n$, while $O_{m\times n}$ is a zero matrix of size $m \times n$. For $x = (x_1, \ldots, x_n)^T$, $\mathsf{diag}(x)$ is a diagonal matrix of the size $n \times n$ with $x_1, x_2, \ldots, x_n$ on its diagonal; $[X]^2$ is $XX^H$, and $\langle X, Y\rangle = \text{trace}(X^H Y)$ for the matrices $X$ and $Y$. Accordingly, the Frobenius norm of $X$ is defined by $||X|| = \sqrt{\text{trace}(X^H X)}$. We also write $\langle X\rangle = \text{trace}(X)$ for notational simplicity. The notation $X \succeq 0$ ($X \succ 0$, resp.) used for the Hermitian symmetric matrix $X$ indicates that it is positive definite (positive semi-definite, resp.). Let us denote the maximal eigenvalue of the Hermitian symmetric matrix $X$ by $\lambda_{\max}(X)$; For a real valued vector $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, $e^{\jmath x}$ is entry-wise understood, i.e. $e^{\jmath x} = (e^{\jmath x_1}, \ldots, e^{\jmath x_n})^T \in \mathbb{C}^n$. For a complex number $x$, $\angle x$ denotes its argument, i.e. $x = e^{\jmath \angle x}$ for $|x| = 1$ and it is fully characterized by $\angle x \in [0, 2\pi)$. Lastly, let us denote the set of circular Gaussian random variables with the zero means and variance $a$ by $\mathcal{C}(0, a)$.

Our mathematical ingredient is the following inequality, which holds true for all matrices $\mathbf{V}$ and $\bar{V}$ of size $n \times m$ and $\mathbf{Y} \succ 0$ and $\bar{Y} \succ 0$ of size $n \times n$ [31]:

$$\begin{aligned}\ln\left|I_n + [\mathbf{V}]^2(\mathbf{Y})^{-1}\right| \geq \ln\left|I_n + [\bar{V}]^2(\bar{Y})^{-1}\right| - \langle [\bar{V}]^2(\bar{Y})^{-1}\rangle \\ + 2\Re\{\langle \bar{V}^H(\bar{Y})^{-1}\mathbf{V}\rangle\} - \left\langle (\bar{Y})^{-1}\right. \\ \left. - (\bar{Y} + [\bar{V}]^2)^{-1}, [\mathbf{V}]^2 + \mathbf{Y}\right\rangle. \quad (1)\end{aligned}$$

Considering both sides of (1) as functions of the variables $(\mathbf{V}, \mathbf{Y})$, they match at $(\bar{V}, \bar{Y})$, i.e. the function defined by the right-hand-side (RHS), which is concave quadratic because $(\bar{Y})^{-1} - (\bar{Y} + [\bar{V}]^2)^{-1} \succeq 0$, provides a tight minorant of the log-determinant function defined by the left-hand-side (LHS) at $(\bar{V}, \bar{Y})$ [32]. The latter is seen as a throughput

function, where $[\mathbf{V}]^2$ plays the role of the covariance of the signal of interest while $\mathbf{Y}$ plays the role of the covariance of the interference-plus-noise signal. As such, the RHS of (1) provides a tight concave quadratic minorant of the throughput functions.

## II. Problem Statements

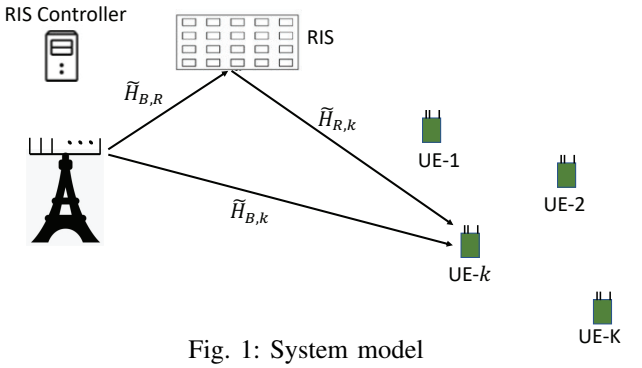


Fig. 1: System model

We consider the RIS-aided communication system illustrated by Fig. 1, where a RIS of $N$ reflecting units supports the downlink (DL) communication from an $N_t$-antenna array BS to $K$ $N_r$-antenna users (UEs) $k \in \mathcal{K} \triangleq \{1, \ldots, K\}$. Since the RIS is typically deployed on the facade of high-rise buildings and the BS is also usually at a certain elevated height [33], it is justified to assume a line-of-sight (LoS) link between the BS and RIS, LoS communication between the RIS and UEs, and NLoS propagation between the BS and UEs. Accordingly, we model the quasi-static and flat-fading channels spanning from the BS and the RIS to UE $k$ and from the BS to the RIS by $\tilde{H}_{B,k} = \sqrt{\beta_{B,k}} H_{B,k} \in \mathbb{C}^{N_r \times N_t}$, $\tilde{H}_{R,k} = \sqrt{\beta_{R,k}} H_{R,k} \in \mathbb{C}^{N_r \times N}$, and $\tilde{H}_{B,R} = \sqrt{\beta_{B,R}} H_{B,R} \in \mathbb{C}^{N \times N_t}$, where $\sqrt{\beta_{B,k}}$, $\sqrt{\beta_{R,k}}$, and $\sqrt{\beta_{B,R}}$ model the path-loss and large-scale fading of the BS-to-UE $k$ link, the RIS-to-UE $k$ link, and the BS-to-RIS link, respectively [5], [34]. Furthermore, $H_{R,k}$ is modelled by Rician fading for representing the LoS channels between the RIS and the UEs [35]. By contrast, $H_{B,k}$ is modelled by Rayleigh fading in the face of non-LoS (NLoS) channels between the BS and the UEs. Like many other papers on RIS-aided communication networks, we assume having perfect channel state information, which can be obtained by channel estimation [4], [8], [36]. Later, we will see the effect of imperfect channel state information (CSI) on the overall system performance in simulation results. The channel matrix of the RIS-aided connection between the BS and UE $k \in \mathcal{K}$ is given by

$$\mathbb{C}^{N_r \times N_t} \ni \mathcal{H}_k(\boldsymbol{\theta}) \triangleq \tilde{H}_{R,k} R_{R,k}^{1/2} \mathsf{diag}(e^{\jmath\boldsymbol{\theta}}) \tilde{H}_{B,R} + \tilde{H}_{B,k} \quad (2)$$

$$= \tilde{H}_{BR,k} \mathsf{diag}(e^{\jmath\boldsymbol{\theta}}) H_{B,R} + \tilde{H}_{B,k}, \quad (3)$$

with $\tilde{H}_{BR,k} \triangleq \sqrt{\beta_{B,R}}\sqrt{\beta_{R,k}} H_{R,k} R_{R,k}^{1/2} \in \mathbb{C}^{N_r \times N}$, where $R_{R,k} \in \mathbb{C}^{N \times N}$ represents the spatial correlation matrix of the RIS elements with respect to user $k$ [5], [37], and $\mathsf{diag}(e^{\jmath\boldsymbol{\theta}})$ in (5) for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^T \in [0, 2\pi)^N$ represents the matrix of PREs. We are interested in quantized PREs having $b$-bit resolution, formulated as:

$$\boldsymbol{\theta}_n \in \mathcal{B} \triangleq \left\{ \nu \frac{2\pi}{2^b}, \nu = 0, 1, \ldots, 2^b - 1 \right\}, \quad (4)$$

for $n \in \mathcal{N} \triangleq \{1, \ldots, N\}$. Let $X \in \mathbb{C}^{N_t \times N_r}$ be the signal to be transmitted from the BS. The signal received at UE $k \in \mathcal{K}$ is

$$y_k = \mathcal{H}_k(\boldsymbol{\theta}) X + n_k, \quad (5)$$

where $n_k \in \mathcal{C}(0, \sigma I_{N_r})$ is the background noise at UE $k$.

We briefly portray a signaling scheme, termed as n-NOMA [17], which includes NOMA and CoSig as a particular case. The UEs are divided into two sets $\mathcal{K}_1 \triangleq \{1, 2, \cdots, K/2\}$ and $\mathcal{K}_2 \triangleq \{K/2+1, \cdots, K\}$, so each UE $k \in \mathcal{K}_1$ is paired with UE $\pi(k) = K/2 + k \in \mathcal{K}_2$ to form a virtual cluster. The widely preferred strategy is to pair users having differentiated channel conditions [38, Sec. II-C], which are determined by their location relative to the BS. Thus the users are ordered based on their geographical distance from the BS. The ordered first user is paired with the ordered $(K/2+1)$-th user. The ordered second user is paired with the ordered $(K/2+2)$-th user and so on.

Under n-NOMA [17], in addition to the information messages $s_k \in \mathcal{C}(0, I_{N_r})$ intended for UE $k \in \mathcal{K}$, there are also information messages $s_{k+K} \in \mathcal{C}(0, I_{N_r})$ intended for UE $\pi(k) \in \mathcal{K}_2$, which are decoded by the paired UEs $k \in \mathcal{K}_1$ and $\pi(k) \in \mathcal{K}_2$.[2]. Each $s_{k'}$ with $k' \in \mathcal{K}_E \triangleq \{1, \ldots, K + K/2\}$ is beamformed by the matrix $\mathbf{W}_{k'} \in \mathbb{C}^{N_t \times N_t}$ to create the signal $X_{k'} = \mathbf{W}_{k'} s_{k'} \in \mathbb{C}^{N_t \times N_r}$, so the transmit signal $X$ in (5) is given by $X = \sum_{k' \in \mathcal{K}_E} \mathbf{W}_{k'} s_{k'}$. The equation (5) becomes

$$y_k = \mathcal{H}_k(\boldsymbol{\theta}) \sum_{k' \in \mathcal{K}_E} \mathbf{W}_{k'} s_{k'} + n_k. \quad (6)$$

Let $\mathbf{W} \triangleq \{\mathbf{W}_{k'}, k' \in \mathcal{K}_E\}$. First, the information $s_{K+k}$ is decoded by the UEs $k \in \mathcal{K}_1$ and $\pi(k) \in \mathcal{K}_2$ with the throughput defined by

$$r_{K+k}(\mathbf{W}, \boldsymbol{\theta}) \triangleq \min\{r_{1,K+k}(\mathbf{W}, \boldsymbol{\theta}), r_{2,K+k}(\mathbf{W}, \boldsymbol{\theta})\}, \quad (7)$$

where $r_{1,K+k}(\mathbf{W}, \boldsymbol{\theta})$ is the throughput of $s_{K+k}$ at UE $k$ defined by

$$r_{1,K+k}(\mathbf{W}, \boldsymbol{\theta}) \triangleq \ln \left| I_{N_r} + [\mathcal{H}_k(\boldsymbol{\theta}) \mathbf{W}_{K+k}]^2 \Lambda_{1,K+k}^{-1}(\mathbf{W}, \boldsymbol{\theta}) \right|, \quad (8)$$

with $\Lambda_{1,K+k}(\mathbf{W}, \boldsymbol{\theta}) \triangleq \sum_{k' \in \mathcal{K}_E \setminus \{K+k\}} [\mathcal{H}_k(\boldsymbol{\theta}) \mathbf{W}_{k'}]^2 + \sigma I_{N_r}$, while $r_{2,K+k}(\mathbf{W}, \boldsymbol{\theta})$ is the throughput of $s_{K+k}$ at UE $\pi(k)$ defined by

$$r_{2,K+k}(\mathbf{W}, \boldsymbol{\theta}) \triangleq \ln \left| I_{N_r} + [\mathcal{H}_{\pi(k)}(\boldsymbol{\theta}) \mathbf{W}_{K+k}]^2 \Lambda_{2,K+k}^{-1}(\mathbf{W}, \boldsymbol{\theta}) \right|, \quad (9)$$

with $\Lambda_{2,K+k}(\mathbf{W}, \boldsymbol{\theta}) \triangleq \sum_{k' \in \mathcal{K}_E \setminus \{K+k\}} [\mathcal{H}_{\pi(k)}(\boldsymbol{\theta}) \mathbf{W}_{k'}]^2 + \sigma I_{N_r}$.

Next, UEs $k \in \mathcal{K}_1$ and $\pi(k) \in \mathcal{K}_2$ subtract $s_{K+k}$ from their received signal to decode $s_k$ and $s_{\pi(k)}$ with the throughput

$$r_\chi(\mathbf{W}, \boldsymbol{\theta}) = \ln \left| I_{N_r} + [\mathcal{H}_\chi(\boldsymbol{\theta}) \mathbf{W}_\chi]^2 \Lambda_\chi^{-1}(\mathbf{W}, \boldsymbol{\theta}) \right|, \quad (10)$$

where $\chi \in \{k, \pi(k)\}$ and $\Lambda_\chi(\mathbf{W}, \boldsymbol{\theta}) = \sum_{k' \in \mathcal{K} \setminus \{\chi, K+k\}} [\mathcal{H}_\chi(\boldsymbol{\theta}) \mathbf{W}_{k'}]^2 + \sigma I_{N_r}$.

[2]This n-NOMA scheme may be regarded as a subclass of Han-Kobayshi signal superposition [39], [40], or rate-splitting based signal superposition [41]

The throughput at UE $\pi(k)$ is

$$r_{\pi(k)}(\mathbf{W},\boldsymbol{\theta}) + \min_{i=1,2} r_{i,K+k}(\mathbf{W},\boldsymbol{\theta}) = \min_{i=1,2} r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta}), \tag{11}$$

for

$$r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta}) \triangleq r_{\pi(k)}(\mathbf{W},\boldsymbol{\theta}) + r_{i,K+k}(\mathbf{W},\boldsymbol{\theta}), i = 1, 2. \tag{12}$$

The problem of maximizing the GM-throughput of n-NOMA is thus formulated as

$$\max_{\mathbf{W},\boldsymbol{\theta}} \; f(\mathbf{W},\boldsymbol{\theta}) \triangleq \left(\prod_{k\in\mathcal{K}_1}\left(r_k(\mathbf{W},\boldsymbol{\theta}) \min_{i=1,2} r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})\right)\right)^{1/K}$$

$$\text{s.t.} \quad (4), \tag{13a}$$

$$\sum_{k'\in\mathcal{K}_E} ||\mathbf{W}_{k'}||^2 \leq P, \tag{13b}$$

where (13b) sets the transmit power constraint, given the power budget $P$.

On one hand, under NOMA, both UE $k$ and UE $\pi(k)$ decode the entire information message intended for UE $\pi(k)$, so NOMA may be viewed as a particular case of the above n-NOMA with

$$\mathbf{W}_{\pi(k)} \equiv 0, k \in \mathcal{K}_1, \tag{14}$$

i.e. the GM-throughput maximization problem (13) corresponding to NOMA is

$$\max_{\mathbf{W},\boldsymbol{\theta}} \left(\prod_{k\in\mathcal{K}_1}\left(r_k(\mathbf{W},\boldsymbol{\theta}) \min_{i=1,2} r_{i,K+k}(\mathbf{W},\boldsymbol{\theta})\right)\right)^{1/K}$$

$$\text{s.t.} \quad (4), (14), (13b), \tag{15}$$

because under (14), $r_{\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ in (11) is zero, and the throughput at UE $\pi(k)$ is defined by $\min_{i=1,2} r_{i,K+k}(\mathbf{W},\boldsymbol{\theta})$. It should be noted that most of the related treatises [13], [14], [25]–[28] enforce the additional constraints of $r_{1,K+k}(\mathbf{W},\boldsymbol{\theta}) \geq r_{2,K+k}(\mathbf{W},\boldsymbol{\theta})$ to have $\min_{i=1,2} r_{i,K+k}(\mathbf{W},\boldsymbol{\theta}) = r_{2,K+k}(\mathbf{W},\boldsymbol{\theta})$, which are computationally intractable and they unnecessarily limit the NOMA feasibility. Our previous contribution [15], [17] showed that this nonsmooth function $\min_{i=1,2} r_{i,K+k}(\mathbf{W},\boldsymbol{\theta})$ can be efficiently handled by convex-solver based computation.

On the other hand, CoSig, under which each information message is decoded by its intended user while treating other messages as interference, may also be interpreted as a particular case of the above n-NOMA along with

$$\mathbf{W}_{K+k} \equiv 0, k \in \mathcal{K}, \tag{16}$$

and the GM-throughput problem (13) corresponding to CoSig is

$$\max_{\mathbf{W},\boldsymbol{\theta}} \; f(\mathbf{W},\boldsymbol{\theta}) \triangleq \left(\prod_{k\in\mathcal{K}} r_k(\mathbf{W},\boldsymbol{\theta})\right)^{1/K} \quad \text{s.t.} \quad (4), (16), (13b), \tag{17}$$

for $r_k(\mathbf{W},\boldsymbol{\theta}) = \ln\left|I_{N_r} + [\mathcal{H}_k(\boldsymbol{\theta})\mathbf{W}_k]^2 \left(\sum_{k'\in\mathcal{K}\setminus\{k\}}[\mathcal{H}_k(\boldsymbol{\theta})\mathbf{W}_{k'}]^2\right)^{-1}\right|$. Our previous contribution [11] considered the particular MISO case ($N_r = 1$) of (17) with the PREs having infinite resolution ($b = \infty$). Again, the motivation of GM-throughput maximization is two-fold. (i) It leads to fair throughput distribution with reasonable ST without enforcing additional computationally-intractable QoS throughput constraints. (ii) It is widely acknowledged that the popular ST maximization and max-min throughput optimization may unfairly allocate excessive resources to the users of favorable and unfavorable users, respectively. This problem is circumvented by GM-throughput optimization at an appealingly low complexity, since there is only a single power constraint [42].

## III. Low-complexity computational solution

Since, the solution of NOMA problem (15) and CoSig problem (17) is a byproduct of n-NOMA problem (13), this section is devoted to addressing the latter. Although the nonsmooth function $\min_{i=1,2} r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ in (13) may not cause difficulty in developing the convex-solver based computational procedures [10], [17], it is an obstacle to deriving the optimal solution of the problem (13) based on closed forms.

By Cauchy's inequality, we have

$$\frac{\sum_{i=1}^2 r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})}{2} \geq \sqrt{\prod_{i=1}^2 r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})}, \tag{18}$$

with equality holding at $r_{1,\pi(k)}(\mathbf{W},\boldsymbol{\theta}) = r_{2,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$. This mean that maximizing the left-hand side (LHS) of (18) leads to unbalanced $r_{1,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ and $r_{2,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ and thus minimizes their minimum, while maximizing the right-hand side (RHS) of (18) leads to balanced $r_{1,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ and $r_{2,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ and thus maximizes their minimum. We hence opt for the RHS of (18) as a surrogate for $\min_{i=1,2} r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})$ in (13). Accordingly, we address (13) via the following surrogate problem

$$\max_{\mathbf{W},\boldsymbol{\theta}} \; f(r(\mathbf{W},\boldsymbol{\theta})) \quad \text{s.t.} \quad (4), (13b), \tag{19}$$

where $f(r(\mathbf{W},\boldsymbol{\theta}))$ is the composite of the function

$$f(r) \triangleq \left(\prod_{k\in\mathcal{K}_1}\left(r_k\sqrt{\prod_{i=1}^2 r_{i,\pi(k)}}\right)\right)^{1/K}, \tag{20}$$

of $K + K/2$ variables and of the mapping, defined as

$$r(\mathbf{W},\boldsymbol{\theta}) \triangleq (r_1(\mathbf{W},\boldsymbol{\theta}), r_{1,\pi(1)}(\mathbf{W},\boldsymbol{\theta}), r_{2,\pi(1)}(\mathbf{W},\boldsymbol{\theta}), \ldots, r_{K/2}(\mathbf{W},\boldsymbol{\theta}), r_{1,\pi(K/2)}(\mathbf{W},\boldsymbol{\theta}), r_{2,\pi(K/2)}(\mathbf{W},\boldsymbol{\theta})). \tag{21}$$

Observe that problem (19) is still very computationally challenging because the objective function in (19) is highly nonlinear and nonconcave while constraint (4) is discrete. We will now develop an iterative procedure for computing (19). In what follows define $\mathbb{R}_+^N = \{(\gamma_1, \ldots, \gamma_N) : \gamma_n > 0, n = 1, \ldots, N\}$ and then

$$\Xi_{\pi(k)} \triangleq \left\{\xi_{\pi(k)} \triangleq (\xi_{1,\pi(k)}, \xi_{2,\pi(k)}) \in \mathbb{R}_+^2 : \prod_{i=1}^2 \xi_{i,\pi(k)} = 1\right\}$$

and

$$\Xi = \left\{(\xi_{\pi(1)}, \ldots, \xi_{\pi(K/2)}) \in \Xi_{\pi(1)} \times \cdots \times \Xi_{\pi(K/2)}\right\},$$

while

$$\Gamma \triangleq \{\gamma \triangleq (\gamma_1, \gamma_{\pi(1)}, \ldots, \gamma_{K/2}, \gamma_{\pi(K/2)}) \in \mathbb{R}_+^K : \prod_{k\in\mathcal{K}_1} \gamma_k \gamma_{\pi(k)} = 1\}.$$

By using the equalities

$$\sqrt{\prod_{i=1}^{2} r_{i,\pi(k)}} = \min_{\xi_{\pi(k)}\in\Xi_{\pi(k)}} \frac{1}{2}\sum_{i=1}^{2} \xi_{i,\pi(k)} r_{i,\pi(k)}$$

and

$$\left(\prod_{k\in\mathcal{K}_1}\left(r_k\sqrt{\prod_{i=1}^{2} r_{i,\pi(k)}}\right)\right)^{1/K} = \min_{\gamma\in\Gamma^{(K)}} \frac{1}{K}\sum_{k\in\mathcal{K}_1}\left(\gamma_k r_k + \gamma_{\pi(k)}\sqrt{\prod_{i=1}^{2} r_{i,\pi(k)})}\right)$$

we can express $f(r)$ in (20) by

$$f(r) = \min_{(\gamma,\xi)\in\Gamma\times\Xi} F(r,\gamma,\xi),$$

with

$$F(r,\gamma,\xi) \triangleq \frac{\sum_{k\in\mathcal{K}_1}\left(\gamma_k r_k + \frac{\gamma_{\pi(k)}}{2}\sum_{i=1}^{2}\xi_{i,\pi(k)} r_{i,\pi(k)}\right)}{K}.$$

As such the problem (19) is equivalent to the following problem of maximin optimization

$$\max_{\mathbf{W},\boldsymbol{\theta}} \min_{(\gamma,\xi)\in\Gamma\times\Xi} F(r(\mathbf{W},\boldsymbol{\theta}),\gamma,\xi) \quad \text{s.t.} \quad (4),(13b). \tag{22}$$

Initialized by its feasible point $(W^{(0)},\theta^{(0)})$, for $\kappa = 1,\ldots,$ we optimize in $(\gamma,\xi)$ to have

$$\xi_{i,\pi(k)}^{(\kappa)} = \frac{\sqrt{\prod_{i'=1}^{2} r_{i',\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})}}{r_{i,\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})}$$

and

$$\gamma_k^{(\kappa)} = \frac{f(r(W^{(\kappa)},\theta^{(\kappa)}))}{r_k(W^{(\kappa)},\theta^{(\kappa)})}, k\in\mathcal{K}_1,$$

while

$$\gamma_{\pi(k)}^{(\kappa)} = \frac{f(r(W^{(\kappa)},\theta^{(\kappa)}))}{\sqrt{\prod_{i=1}^{2} r_{i,\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})}}, k\in\mathcal{K}_1.$$

For $f^{(\kappa)}(\mathbf{W},\boldsymbol{\theta}) \triangleq F(r(\mathbf{W},\boldsymbol{\theta}),\gamma^{(\kappa)},\xi^{(\kappa)})$ which is

$$\sum_{k\in\mathcal{K}_1}\left[\gamma_k^{(\kappa)} r_k(\mathbf{W},\boldsymbol{\theta}) + \sum_{i=1}^{2}\gamma_{i,\pi(k)}^{(\kappa)} r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})\right],$$

with $\gamma_{i,\pi(k)}^{(\kappa)} = \frac{1}{2}\gamma_{\pi(k)}^{(\kappa)}\xi_{i,\pi(k)}^{(\kappa)} = \frac{f(r(W^{(\kappa)},\theta^{(\kappa)}))}{2r_{i,\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})}$, $i=1,2$, we iterate $(W^{(\kappa+1)},\theta^{(\kappa+1)})$ by solving the problem of mixed continuous-discrete optimization

$$\max_{\mathbf{W}\,\boldsymbol{\theta}\in\mathcal{B}^N} f^{(\kappa)}(\mathbf{W},\boldsymbol{\theta}) \quad \text{s.t.} \quad (4),(13b). \tag{23}$$

In what follows, we use the following definition:

$$H_k^{(\kappa)} \triangleq \mathcal{H}_k(\theta^{(\kappa)}), k\in\mathcal{K}; H_{1,k}^{(\kappa)} \equiv H_k^{(\kappa)} \;\&\; H_{2,k}^{(\kappa)} \equiv H_{\pi(k)}^{(\kappa)}. \tag{24}$$

### A. Beamforming iteration

We seek $W^{(\kappa+1)}$ such that

$$f^{(\kappa)}(W^{(\kappa+1)},\theta^{(\kappa)}) > f^{(\kappa)}(W^{(\kappa)},\theta^{(\kappa)}). \tag{25}$$

Define $r_{b,k}^{(\kappa)}(\mathbf{W}) \triangleq r_k(\mathbf{W},\theta^{(\kappa)})$, $k\in\mathcal{K}$, and $r_{b,i,\pi(k)}^{(\kappa)}(\mathbf{W}) \triangleq r_{i,\pi(k)}(\mathbf{W},\theta^{(\kappa)})$, and $r_{b,i,K+k}^{(\kappa)}(\mathbf{W}) \triangleq r_{i,K+k}(\mathbf{W},\theta^{(\kappa)})$, $i=1,2$; $k\in\mathcal{K}_1$.

Then for $k\in\mathcal{K}$, applying the inequality (1) yields the following tight concave quadratic minorant of $r_{b,k}^{(\kappa)}(\mathbf{W})$ at $W^{(\kappa)}$:

$$\tilde{r}_{b,k}^{(\kappa)}(\mathbf{W}) \triangleq a_k^{(\kappa)} + 2\Re\{\langle B_k^{(\kappa)},\mathbf{W}_k\rangle\} - \sum_{k'\in\mathcal{K}_E}\langle(\mathbf{W}_{k'})^H\mathcal{C}_{k,k'}^{(\kappa)}\mathbf{W}_{k'}\rangle, \tag{26}$$

where $a_k^{(\kappa)} \triangleq r_{b,k}^{(\kappa)}(W^{(\kappa)}) - \langle[H_k^{(\kappa)}W_k^{(\kappa)}]^2(\Lambda_k^{(\kappa)})^{-1}\rangle - \langle C_k^{(\kappa)}\rangle\sigma$, $B_k^{(\kappa)} \triangleq (H_k^{(\kappa)})^H(\Lambda_k^{(\kappa)})^{-1}H_k^{(\kappa)}W_k^{(\kappa)}$, and

$$\mathcal{C}_{k,k'}^{(\kappa)} \triangleq \begin{cases} O_{N_t\times N_t} & \text{for } k' = K+k \\ (H_k^{(\kappa)})^H C_k^{(\kappa)} H_k^{(\kappa)} & \text{otherwise,}\end{cases}$$

with $\Lambda_k^{(\kappa)} \triangleq \Lambda_k(W^{(\kappa)},\theta^{(\kappa)})$, and $C_k^{(\kappa)} \triangleq (\Lambda_k^{(\kappa)})^{-1} - (\Lambda_k^{(\kappa)} + [H_k^{(\kappa)}W_k^{(\kappa)}]^2)^{-1}$.

Analogously, a tight concave quadratic minorant of $r_{b,i,K+k}^{(\kappa)}(\mathbf{W})$ at $W^{(\kappa)}$ is

$$\tilde{r}_{b,i,K+k}^{(\kappa)}(\mathbf{W}) \triangleq a_{i,K+k}^{(\kappa)} + 2\Re\{\langle B_{i,K+k}^{(\kappa)},\mathbf{W}_{K+k}\rangle\} - \sum_{k'\in\mathcal{K}_E}\langle(\mathbf{W}_{k'})^H\mathcal{C}_{i,K+k}^{(\kappa)}\mathbf{W}_{k'}\rangle, \tag{27}$$

where $a_{i,K+k}^{(\kappa)} \triangleq r_{b,i,K+k}^{(\kappa)}(W^{(\kappa)}) - \langle[H_{i,k}^{(\kappa)}W_{K+k}^{(\kappa)}]^2(\Lambda_{i,K+k}^{(\kappa)})^{-1}\rangle - \sigma\langle C_{i,K+k}^{(\kappa)}\rangle$, $B_{1,K+k}^{(\kappa)} \triangleq (H_{i,k}^{(\kappa)})^H(\Lambda_{i,K+k}^{(\kappa)})^{-1}H_{i,k}^{(\kappa)}W_{K+k}^{(\kappa)})$, and $\mathcal{C}_{i,K+k}^{(\kappa)} \triangleq (H_{i,k}^{(\kappa)})^H C_{i,K+k}^{(\kappa)}H_k^{(\kappa)}$, with $\Lambda_{i,K+k}^{(\kappa)} \triangleq \Lambda_{i,K+k}(W^{(\kappa)},\theta^{(\kappa)})$, and $C_{i,K+k}^{(\kappa)} \triangleq (\Lambda_{i,K+k}^{(\kappa)})^{-1} - (\Lambda_{1,K+k}^{(\kappa)} + [H_{i,k}^{(\kappa)}W_{K+k}^{(\kappa)}]^2)^{-1}$, $i=1,2$.

Then, we have the following tight concave quadratic minorant of $r_{b,i,\pi(k)}^{(\kappa)}(\mathbf{W})$ at $W^{(\kappa)}$:

$$\begin{aligned}\tilde{r}_{b,i,\pi(k)}^{(\kappa)}(\mathbf{W}) &\triangleq \tilde{r}_{b,\pi(k)}^{(\kappa)}(\mathbf{W}) + \tilde{r}_{b,i,K+k}^{(\kappa)}(\mathbf{W}) \\ &= a_{i,\pi(k)}^{(\kappa)} + 2\Re\{\langle B_{\pi(k)}^{(\kappa)},\mathbf{W}_{\pi(k)}\rangle\} \\ &\quad + 2\Re\{\langle B_{i,K+k}^{(\kappa)}\mathbf{W}_{K+k}\rangle\} \\ &\quad - \sum_{k'\in\mathcal{K}_E}\langle(\mathbf{W}_{k'})^H\mathcal{C}_{i,\pi(k),k'}^{(\kappa)}\mathbf{W}_{k'}\rangle,\end{aligned} \tag{28}$$

where $a_{i,\pi(k)}^{(\kappa)} \triangleq a_{\pi(k)}^{(\kappa)} + a_{i,K+k}^{(\kappa)}$, and $\mathcal{C}_{i,\pi(k),k'}^{(\kappa)} = \mathcal{C}_{\pi(k),k'}^{(\kappa)} + \mathcal{C}_{i,K+k}^{(\kappa)}$, $i=1,2$.

In summary, a tight concave quadratic minorant of $f^{(\kappa)}(\mathbf{W},\theta^{(\kappa)}) \geq \tilde{f}_b^{(\kappa)}(\mathbf{W})$ is

$$\begin{aligned}\tilde{f}_b^{(\kappa)}(\mathbf{W}) &\triangleq \sum_{k\in\mathcal{K}_1}\left[\gamma_k^{(\kappa)}\tilde{r}_{b,k}^{(\kappa)}(\mathbf{W}) + \sum_{i=1}^{2}\gamma_{i,\pi(k)}^{(\kappa)}\tilde{r}_{b,i,\pi(k)}^{(\kappa)}(\mathbf{W})\right] \\ &= a^{(\kappa)} + 2\sum_{k\in\mathcal{K}_E}\Re\{\langle\tilde{B}_k^{(\kappa)},\mathbf{W}_k\rangle\}\end{aligned}$$

$$- \sum_{k\in\mathcal{K}_E} \langle (\mathbf{W}_k)^H \Psi_k^{(\kappa)} \mathbf{W}_k \rangle, \quad (29)$$

for $a^{(\kappa)} \triangleq \sum_{k\in\mathcal{K}_1} \left( \gamma_k^{(\kappa)} a_k^{(\kappa)} + \sum_{i=1}^2 \gamma_{i,\pi(k)}^{(\kappa)} a_{i,\pi(k)}^{(\kappa)} \right)$,[3]

$$\tilde{B}_k^{(\kappa)} \triangleq \begin{cases} \gamma_k^{(\kappa)} B_k^{(\kappa)}, & k \in \mathcal{K}_1 \\ (\sum_{i=1}^2 \gamma_{1,\pi^{-1}(k)}^{(\kappa)}) B_{\pi^{-1}(k)}^{(\kappa)}, & k \in \mathcal{K}_2 \\ \sum_{i=1}^2 \gamma_{i,\pi(k-K)}^{(\kappa)} B_{k,k-K}^{(\kappa)}, & k > K, \end{cases}$$

and $\Psi_k^{(\kappa)} \triangleq \sum_{k'\in\mathcal{K}_1} \left( \gamma_{k'}^{(\kappa)} \mathcal{C}_{k',k}^{(\kappa)} + \sum_{i=1}^2 \gamma_{i,\pi(k')}^{(\kappa)} \mathcal{C}_{i,\pi(k'),k}^{(\kappa)}, k \in \mathcal{K}_E$.

We solve the following convex quadratic problem of tight minorant maximization for (23) to generate $W^{(\kappa+1)}$

$$\max_{\mathbf{W}} \ \tilde{f}_b^{(\kappa)}(\mathbf{W}) \quad \text{s.t.} \quad (13b), \quad (30)$$

which admits the closed-form solution of

$$W_k^{(\kappa+1)} = \begin{cases} (\Psi_k^{(\kappa)})^{-1} \tilde{B}_k^{(\kappa)}, & \text{if } \Xi^{(\kappa)} \le P \\ (\Psi_k^{(\kappa)} + \mu^{(\kappa)} I_{N_t})^{-1} \tilde{B}_k^{(\kappa)}, & \text{otherwise,} \end{cases} \quad (31)$$

where $\Xi^{(\kappa)} \triangleq \sum_{k\in\mathcal{K}_E} ||(\Psi_k^{(\kappa)})^{-1} \tilde{B}_k^{(\kappa)}||^2$ and $\mu^{(\kappa)} > 0$ is found by bisection, such that $\sum_{k\in\mathcal{K}_E} ||(\Psi_k^{(\kappa)} + \mu^{(\kappa)} I_{N_t})^{-1} \tilde{B}_k^{(\kappa)}||^2 = P$.

As $W^{(\kappa+1)}$ and $W^{(\kappa)}$ are the optimal solution and a feasible point for (30), it follows that $\tilde{f}_b^{(\kappa)}(W^{(\kappa+1)}) > \tilde{f}_b^{(\kappa)}(W^{(\kappa)})$ as far as $\tilde{f}_b^{(\kappa)}(W^{(\kappa+1)}) \neq \tilde{f}_b^{(\kappa)}(W^{(\kappa)})$. We then have $f^{(\kappa)}(W^{(\kappa+1)}, \theta^{(\kappa)}) \ge \tilde{f}_b^{(\kappa)}(W^{(\kappa+1)}) > \tilde{f}_b^{(\kappa)}(W^{(\kappa)}) = f^{(\kappa)}(W^{(\kappa)}, \theta^{(\kappa)})$, verifying (25).

### *B. PREs iteration*

Similarly to (25), now we seek $\theta^{(\kappa+1)}$, such that

$$f^{(\kappa)}(W^{(\kappa+1)}, \theta^{(\kappa+1)}) > f^{(\kappa)}(W^{(\kappa+1)}, \theta^{(\kappa)}). \quad (32)$$

Define $r_{p,k}^{(\kappa)}(\boldsymbol{\theta}) \triangleq r_k(W^{(\kappa+1)}, \boldsymbol{\theta})$, $k \in \mathcal{K}$, and $r_{p,i,\pi(k)}^{(\kappa)}(\boldsymbol{\theta}) \triangleq r_{i,\pi(k)}(W^{(\kappa+1)}, \boldsymbol{\theta})$, and $r_{p,i,K+k}^{(\kappa)}(\boldsymbol{\theta}) \triangleq r_{i,K+k}(W^{(\kappa+1)}, \boldsymbol{\theta})$, $i = 1, 2$; $k \in \mathcal{K}_1$.

For $k \in \mathcal{K}$, using the inequality (1) yields the following minorant of $r_{p,k}^{(\kappa)}(\boldsymbol{\theta})$ at $\theta^{(\kappa)}$:

$$\tilde{r}_{p,k}^{(\kappa)}(\boldsymbol{\theta}) \triangleq \tilde{a}_k^{(\kappa)} + 2\Re\{\langle (H_k^{(\kappa)} W_k^{(\kappa+1)})^H (\widetilde{\Lambda}_k^{(\kappa)})^{-1} \mathcal{H}_k(\boldsymbol{\theta}) \times W_k^{(\kappa+1)} \rangle\} - \langle \tilde{C}_k^{(\kappa)}, \mathcal{H}_k(\boldsymbol{\theta}) \mathcal{W}_k^{(\kappa+1)} \mathcal{H}_k^H(\boldsymbol{\theta}) \rangle, \quad (33)$$

with $\tilde{a}_k^{(\kappa)} \triangleq r_{p,k}^{(\kappa)}(\theta^{(\kappa)}) - \langle [H_k^{(\kappa)} W_k^{(\kappa+1)}]^2 (\widetilde{\Lambda}_k^{(\kappa)})^{-1} \rangle - \sigma \langle \tilde{C}_k^{(\kappa)} \rangle$, $\widetilde{\Lambda}_k^{(\kappa)} \triangleq \Lambda_k(W^{(\kappa+1)}, \theta^{(\kappa)})$, $0 \preceq \tilde{C}_k^{(\kappa)} \triangleq (\widetilde{\Lambda}_k^{(\kappa)})^{-1} - (\widetilde{\Lambda}_k^{(\kappa)} + [H_k^{(\kappa)} W_k^{(\kappa+1)}]^2)^{-1}$, and $\mathcal{W}_k^{(\kappa+1)} \triangleq \sum_{k'\in\mathcal{K}_E\setminus\{K+k\}} [W_{k'}^{(\kappa+1)}]^2$.

We may write $\mathrm{diag}(e^{\jmath\boldsymbol{\theta}}) = \sum_{n=1}^N e^{\jmath\boldsymbol{\theta}_n} \Upsilon_n$, where $\Upsilon_n$ is the matrix of size $N \times N$ having only zero entries, except for its $(n,n)$-entry, which is 1. Then the matrix $\mathcal{H}_k(\boldsymbol{\theta})$ defined by (3) is represented by $\mathcal{H}_k(\boldsymbol{\theta}) = \sum_{n=1}^N e^{\jmath\boldsymbol{\theta}_n} \mathcal{H}_{k,n} + \tilde{H}_{B,k}$ with $\mathcal{H}_{k,n} \triangleq \tilde{H}_{BR,k} \Upsilon_n H_{B,R}$. Therefore,

$$\langle (H_k^{(\kappa)} W_k^{(\kappa+1)})^H (\widetilde{\Lambda}_k^{(\kappa)})^{-1} \mathcal{H}_k(\boldsymbol{\theta}) W_k^{(\kappa+1)} \rangle = \alpha_{k,1}^{(\kappa)} + \sum_{n=1}^N \tilde{b}_{k,1}^{(\kappa)}(n) e^{\jmath\boldsymbol{\theta}_n}, \quad (34)$$

with[4] $\alpha_{k,1}^{(\kappa)} \triangleq \langle (H_k^{(\kappa)} W_k^{(\kappa+1)})^H (\widetilde{\Lambda}_k^{(\kappa)})^{-1} \tilde{H}_{B,k} W_k^{(\kappa+1)} \rangle$, $\tilde{b}_{k,1}^{(\kappa)}(n) = \langle (H_k^{(\kappa)} W_k^{(\kappa+1)})^H (\widetilde{\Lambda}_k^{(\kappa)})^{-1} \mathcal{H}_{k,n} W_k^{(\kappa+1)} \rangle$, $n \in \mathcal{N}$. Furthermore,

$$\langle \tilde{C}_k^{(\kappa)}, \mathcal{H}_k(\boldsymbol{\theta}) \mathcal{W}_k^{(\kappa+1)} \mathcal{H}_k^H(\boldsymbol{\theta}) \rangle = \alpha_{k,2}^{(\kappa)} + 2\Re\{\tilde{b}_{k,2}^{(\kappa)}(n) e^{\jmath\boldsymbol{\theta}_n}\} + (e^{\jmath\boldsymbol{\theta}})^H \Phi_k^{(\kappa+1)} e^{\jmath\boldsymbol{\theta}}, \quad (35)$$

with $\alpha_{k,2}^{(\kappa)} \triangleq \langle (\tilde{H}_{B,k})^H \tilde{C}_k^{(\kappa)} \tilde{H}_{B,k} \mathcal{W}_k^{(\kappa+1)} \rangle$, and $\tilde{b}_{k,2}^{(\kappa)}(n) = \langle (\tilde{H}_{B,k})^H \tilde{C}_k^{(\kappa)} \mathcal{H}_{k,n} \mathcal{W}_k^{(\kappa+1)} \rangle$, $n = 1, \ldots, N$, and $\Phi_k^{(\kappa+1)}(n', n) \triangleq \langle \mathcal{H}_{k,n'}^H \tilde{C}_k^{(\kappa)} \mathcal{H}_{k,n} \mathcal{W}_k^{(\kappa+1)} \rangle$, $(n', n) \in \mathcal{N} \times \mathcal{N}$.

Based on (34), and (35), we obtain

$$\tilde{r}_{p,k}^{(\kappa)}(\boldsymbol{\theta}) = \tilde{a}_k^{(\kappa+1)} + 2\Re\Big\{ \sum_{n=1}^N \tilde{b}_k^{(\kappa+1)}(n) e^{\jmath\boldsymbol{\theta}_n} \Big\} - (e^{\jmath\boldsymbol{\theta}})^H \Phi_k^{(\kappa+1)} e^{\jmath\boldsymbol{\theta}}, \quad (36)$$

with $\tilde{a}_k^{(\kappa+1)} \triangleq \tilde{a}_k^{(\kappa)} + 2\Re\{\alpha_{k,1}^{(\kappa)}\} - \alpha_{k,2}^{(\kappa)}$, $\tilde{b}_k^{(\kappa+1)}(n) \triangleq \tilde{b}_{k,1}^{(\kappa)}(n) - \tilde{b}_{k,2}^{(\kappa)}(n)$, $n \in \mathcal{N}$.

Analogously, under the definition (24), and $\tilde{H}_{1,B,k} \triangleq \tilde{H}_{B,k}$, $\tilde{H}_{2,B,k} \triangleq \tilde{H}_{B,\pi(k)}$, while $\mathcal{H}_{1,k,n} \triangleq \mathcal{H}_{k,n}$ and $\mathcal{H}_{2,k,n} \triangleq \mathcal{H}_{\pi(k),n}$, we obtain the following minorant of $r_{p,i,K+k}^{(\kappa)}(\boldsymbol{\theta})$ at $\theta^{(\kappa)}$

$$\tilde{r}_{p,i,K+k}^{(\kappa)}(\boldsymbol{\theta}) \triangleq \tilde{a}_{1,K+k}^{(\kappa+1)} + 2\Re\{\sum_{n=1}^N \tilde{b}_{1,K+k}^{(\kappa+1)}(n) e^{\jmath\boldsymbol{\theta}_n}\} - (e^{\jmath\boldsymbol{\theta}})^H \Phi_{1,K+k}^{(\kappa+1)} e^{\jmath\boldsymbol{\theta}}, \quad (37)$$

where $(i)$ $\tilde{a}_{i,K+k}^{(\kappa+1)} \triangleq \tilde{a}_{i,K+k}^{(\kappa)} + 2\Re\{\alpha_{i,K+k,1}^{(\kappa)}\} - \alpha_{i,K+k,2}^{(\kappa)}$ with $\tilde{a}_{i,K+k}^{(\kappa)} \triangleq r_{p,i,K+k}^{(\kappa)}(\theta^{(\kappa)}) - \langle [H_{i,k}^{(\kappa)} W_{K+k}^{(\kappa+1)}]^2 (\widetilde{\Lambda}_{i,K+k}^{(\kappa)})^{-1} \rangle - \sigma \langle \tilde{C}_{i,K+k}^{(\kappa)} \rangle$ with $\widetilde{\Lambda}_{i,K+k}^{(\kappa)} \triangleq \Lambda_{i,K+k}(W^{(\kappa+1)}, \theta^{(\kappa)})$ and $0 \preceq \tilde{C}_{i,K+k}^{(\kappa)} \triangleq (\widetilde{\Lambda}_{i,K+k}^{(\kappa)})^{-1} - (\widetilde{\Lambda}_{i,K+k}^{(\kappa)} + [H_{i,k}^{(\kappa)} W_{K+k}^{(\kappa+1)}]^2)^{-1}$, and $\alpha_{i,K+k,1}^{(\kappa)} \triangleq \langle (H_{i,k}^{(\kappa)} W_{K+k}^{(\kappa+1)})^H (\widetilde{\Lambda}_{1,K+k}^{(\kappa)})^{-1} \tilde{H}_{i,B,k} W_{K+k}^{(\kappa+1)} \rangle$, and $\alpha_{i,K+k,2}^{(\kappa)} \triangleq \langle (\tilde{H}_{i,B,k})^H \tilde{C}_{i,K+k}^{(\kappa)} \tilde{H}_{i,B,k} \mathcal{W}_{K+k}^{(\kappa+1)} \rangle$; $(ii)$ $\tilde{b}_{i,K+k}^{(\kappa+1)}(n) \triangleq \tilde{b}_{i,K+k,1}^{(\kappa)}(n) - \tilde{b}_{i,K+k,2}^{(\kappa)}(n)$, $n \in \mathcal{N}$, with $\tilde{b}_{i,K+k,1}^{(\kappa)}(n) = \langle (H_{i,k}^{(\kappa)} W_{K+k}^{(\kappa+1)})^H (\widetilde{\Lambda}_{i,K+k}^{(\kappa)})^{-1} \mathcal{H}_{i,k,n} W_{K+k}^{(\kappa+1)} \rangle$, and $\tilde{b}_{i,K+k,2}^{(\kappa)}(n) \triangleq \langle (\tilde{H}_{i,B,k})^H \tilde{C}_{i,K+k}^{(\kappa)} \mathcal{H}_{i,k,n} \mathcal{W}_{K+k}^{(\kappa+1)} \rangle$ with $\mathcal{W}_{K+k}^{(\kappa+1)} \triangleq \sum_{k'\in\mathcal{K}_E} [W_{k'}^{(\kappa+1)}]^2$, $n \in \mathcal{N}$; and $(iii)$ $\Phi_{i,K+k}^{(\kappa+1)}(n, n') \triangleq \langle \mathcal{H}_{i,k,n'}^H \tilde{C}_{i,K+k}^{(\kappa)} \mathcal{H}_{i,k,n} \mathcal{W}_{K+k}^{(\kappa+1)} \rangle$, $(n', n) \in \mathcal{N} \times \mathcal{N}$.

Then a minorant of $r_{p,i,\pi(k)}^{(\kappa)}(\boldsymbol{\theta})$ at $\theta^{(\kappa)}$ is

$$\begin{aligned} \tilde{r}_{p,i,\pi(k)}^{(\kappa)}(\boldsymbol{\theta}) &\triangleq \tilde{r}_{p,\pi(k)}^{(\kappa)}(\boldsymbol{\theta}) + \tilde{r}_{p,1,K+k}^{(\kappa)}(\boldsymbol{\theta}) \\ &= \tilde{a}_{i,\pi(k)}^{(\kappa+1)} + 2\Re\left\{ \sum_{n=1}^N \tilde{b}_{i,\pi(k)}^{(\kappa+1)}(n) e^{\jmath\boldsymbol{\theta}_n} \right\} \\ &\quad - (e^{\jmath\boldsymbol{\theta}})^H \Phi_{i,\pi(k)}^{(\kappa+1)} e^{\jmath\boldsymbol{\theta}}, \end{aligned} \quad (38)$$

where $\tilde{a}_{i,\pi(k)}^{(\kappa+1)} \triangleq \tilde{a}_{\pi(k)}^{(\kappa+1)} + \tilde{a}_{i,K+k}^{(\kappa+1)}$, $\tilde{b}_{i,\pi(k)}^{(\kappa+1)} \triangleq \tilde{b}_{\pi(k)}^{(\kappa+1)} + \tilde{b}_{i,K+k}^{(\kappa+1)}$, and $\Phi_{i,\pi(k)}^{(\kappa+1)} \triangleq \Phi_{\pi(k)}^{(\kappa+1)} + \Phi_{i,K+k}^{(\kappa+1)}$, $i = 1, 2$.

[3] For $k \in \mathcal{K}_2$, $\pi^{-1}(k)$ is $k' \in \mathcal{K}_1$ such that $\pi(k') = k$

[4] In what follows $b(i)$ is the $i$-th entry of $b$ and $[A](i,i)$ is the $i$-th diagonal entry of $A$, and $[A]^*(i,i)$ is the complex conjugate of $[A](i,i)$

In summary, a minorant of $f^{(\kappa)}(W^{(\kappa+1)},\boldsymbol{\theta})$ at $\theta^{(\kappa)}$ is

$$\begin{aligned}\tilde{f}_p^{(\kappa)}(\boldsymbol{\theta}) &\triangleq \sum_{k\in\mathcal{K}_1}\Big[\gamma_k^{(\kappa)}\tilde{r}_{p,k}^{(\kappa)}(\boldsymbol{\theta})+\sum_{i=1}^{2}\gamma_{i,\pi(k)}^{(\kappa)}\tilde{r}_{p,i,\pi(k)}^{(\kappa)}(\boldsymbol{\theta})\Big]\\ &= \tilde{a}_p^{(\kappa+1)}+2\Re\{\sum_{n=1}^{N}\tilde{b}_p^{(\kappa+1)}(n)e^{\jmath\boldsymbol{\theta}_n}\}\\ &\quad -(e^{\jmath\boldsymbol{\theta}})^H\Phi^{(\kappa+1)}e^{\jmath\boldsymbol{\theta}}, \qquad (39)\end{aligned}$$

for $\tilde{a}_p^{(\kappa+1)} \triangleq \sum_{k\in\mathcal{K}_1}\big(\gamma_k^{(\kappa)}\tilde{a}_k^{(\kappa+1)}+\sum_{i=1}^{2}\gamma_{i,\pi(k)}^{(\kappa)}\tilde{a}_{i,\pi(k)}^{(\kappa+1)}\big)$, $\tilde{b}_p^{(\kappa+1)} \triangleq \sum_{k\in\mathcal{K}_1}\big(\gamma_k^{(\kappa)}\tilde{b}_k^{(\kappa+1)}+\sum_{i=1}^{2}\gamma_{i,\pi(k)}^{(\kappa)}\tilde{b}_{i,\pi(k)}^{(\kappa+1)}\big)$, and $\Phi^{(\kappa+1)} \triangleq \sum_{k\in\mathcal{K}_1}\big(\gamma_k^{(\kappa)}\Phi_k^{(\kappa+1)}+\sum_{i=1}^{2}\gamma_{i,\pi(k)}^{(\kappa)}\Phi_{i,\pi(k)}^{(\kappa+1)}\big)$.

Furthermore,

$$\begin{aligned}\tilde{f}_p^{(\kappa)}(\boldsymbol{\theta}) &\geq \tilde{a}_p^{(\kappa+1)}+2\Re\Big\{\sum_{n=1}^{N}\Big(\tilde{b}_p^{(\kappa+1)}(n)-\sum_{n'=1}^{N}e^{-\jmath\theta_{n'}^{(\kappa)}}\\ &\quad\times\Phi^{(\kappa+1)}(n',n)+\lambda_{\max}(\Phi^{(\kappa+1)})e^{-\jmath\theta_n^{(\kappa)}}\Big)e^{\jmath\boldsymbol{\theta}_n}\Big\}\\ &\quad-(e^{\jmath\theta^{(\kappa)}})^H\Phi^{(\kappa+1)}e^{\jmath\theta^{(\kappa)}}-2\lambda_{\max}(\Phi^{(\kappa+1)})N \qquad (40)\\ &\triangleq \tilde{\tilde{f}}_p^{(\kappa)}(\boldsymbol{\theta}). \qquad (41)\end{aligned}$$

The function $\tilde{\tilde{f}}_p^{(\kappa)}(\boldsymbol{\theta})$ is still a tight minorant of $f^{(\kappa)}(W^{(\kappa+1)},\boldsymbol{\theta})$ at $\theta^{(\kappa)}$, because we have $\tilde{\tilde{f}}_p^{(\kappa)}(\theta^{(\kappa)})=f^{(\kappa)}(W^{(\kappa+1)},\theta^{(\kappa)})$ We thus solve the following discrete problem of tight minorant maximization for (23) to generate $\theta^{(\kappa+1)}$:

$$\max_{\boldsymbol{\theta}\in\mathcal{B}^N}\ \tilde{\tilde{f}}_p^{(\kappa)}(\boldsymbol{\theta}), \qquad (42)$$

which admits the closed-form solution of

$$\begin{aligned}\theta_n^{(\kappa+1)} &= 2\pi-\Big\lfloor\angle\Big(\tilde{b}_p^{(\kappa+1)}(n)-\sum_{m=1}^{N}e^{-\jmath\theta_m^{(\kappa)}}\Phi^{(\kappa+1)}(m,n)\\ &\quad+\lambda_{\max}(\Phi^{(\kappa+1)})e^{-\jmath\theta_n^{(\kappa)}}\Big)\Big\rceil_b, \quad n\in\mathcal{N}, \qquad (43)\end{aligned}$$

where $\lfloor\alpha\rceil_b$ represents the projection of $\alpha\in[0,2\pi]$ into $\mathcal{B}$ defined by

$$\lfloor\alpha\rceil_b=\nu_\alpha\frac{2\pi}{2^b}, \qquad (44)$$

with

$$\nu_\alpha \triangleq \arg\min_{\nu=0,1,\ldots,2^b}\left|\nu\frac{2\pi}{2^b}-\alpha\right|, \qquad (45)$$

which can be readily found because we have $\nu_\alpha\in\{\nu,\nu+1\}$ for $\alpha\in[\nu\frac{2\pi}{2^b},(\nu+1)\frac{2\pi}{2^b}]$. We also reset $\nu_\alpha=0$, when the optimal solution of (45) is $2^b$.

It follows from (40) that $f^{(\kappa)}(W^{(\kappa+1)},\theta^{(\kappa+1)}) \geq \tilde{\tilde{f}}_p^{(\kappa)}(\theta^{(\kappa+1)}) > \tilde{\tilde{f}}_p^{(\kappa)}(\theta^{(\kappa)}) = f^{(\kappa)}(W^{(\kappa+1)},\theta^{(\kappa)})$, confirming (32).

### C. GM-throughput optimization algorithm

Algorithm 1 provides the pseudo-code for the procedure proposed for computing (19). It can be observed from (25) and (32), which are proved at the end of Sections III-A and III-B, respectively, that the objective function in (23) improves after each iteration. Since (23) and (19) share the same first order optimality condition, this implies that Algorithm 1, which provides a procedure for computing (19), converges to a local solution satisfying the first order optimality condition. Interestingly, it has been consistently observed from our simulations that

$$f(r(W^{(\kappa+1)},\theta^{(\kappa+1)})) > f(r(W^{(\kappa)},\theta^{(\kappa)})), \qquad (46)$$

i.e., Algorithm 1 provides a path-following procedure for computing (19) and generates a sequence of improved feasible points to converge at least to a locally optimal solution of the problem (19) [43].

**Algorithm 1** GM-throughput maximization algorithm

1: **Initialization:** Set $\kappa=0$. Randomly generate $(W^{(0)},\theta^{(0)})$ satisfying the constraints (4) and (13b).
2: **Repeat until convergence of the objective in** (19)**:** Generate $W^{(\kappa+1)}$ by (31) and $\theta^{(\kappa+1)}$ by (43). Reset $\kappa\leftarrow\kappa+1$.
3: **Output** $(W^{(\kappa)},\theta^{(\kappa)})$ and rates $r_k(W^{(\kappa)},\theta^{(\kappa)})$, $k\in\mathcal{K}$ with their achieved GM $\Big[\prod_{k\in\mathcal{K}_1}\Big(r_k(W^{(\kappa)},\theta^{(\kappa)})\min_{i=1,2}\big\{r_{\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})+r_{i,K+k}(W^{(\kappa)},\theta^{(\kappa)})\big\}\Big)\Big]^{1/K}$.

### D. Sum throughput optimization

It is straightforward to adjust Algorithm 1 to address the problem of ST maximization:

$$\max_{\mathbf{W},\boldsymbol{\theta}\in\mathcal{B}^N}\sum_{k\in\mathcal{K}_1}\Big(r_k(\mathbf{W},\boldsymbol{\theta})+\min_{i=1,2}r_{i,\pi(k)}(\mathbf{W},\boldsymbol{\theta})\Big) \quad \text{s.t.} \quad (13b), \qquad (47)$$

via the following surrogate problem

$$\begin{aligned}\max_{\mathbf{W},\boldsymbol{\theta}\in\mathcal{B}^N}\sum_{k\in\mathcal{K}_1}&\Big(r_k(\mathbf{W},\boldsymbol{\theta})\\ &+\sqrt{r_{1,\pi(k)}(\mathbf{W},\boldsymbol{\theta})r_{2,\pi(k)}(\mathbf{W},\boldsymbol{\theta})}\Big) \quad \text{s.t.} \quad (13b). \qquad (48)\end{aligned}$$

Algorithm 1 is ready to compute (48) with $\gamma_k^{(\kappa)}$ and $\gamma_{i,\pi(k)}^{(\kappa)}$ for $k\in\mathcal{K}_1$ in (23) adjusted to $\gamma_k^{(\kappa)}\equiv 1$, and $\gamma_{i,\pi(k)}^{(\kappa)}=\frac{\sqrt{r_{1,\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})r_{2,\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})}}{2r_{i,\pi(k)}(W^{(\kappa)},\theta^{(\kappa)})}$, $i=1,2$.

## IV. GM-throughput maximization under individual transmit-antenna power constraints

Instead of the sum power constraint (13b), which is convex quadratic, we now consider individual per-TA equal-power constraints to support low-cost amplification at each TA:

$$\sum_{k'\in\mathcal{K}_E}\sum_{j=1}^{N_r}|\mathbf{W}_{k'}(i,j)|^2=P/N_t, i=1,\ldots,N_t, \qquad (49)$$

which are not convex, i.e. the objective of this section is to solve the following problem:

$$\max_{\mathbf{W},\boldsymbol{\theta}}\ f(\mathbf{W},\boldsymbol{\theta}) \qquad \text{s.t.} \quad (4),(49). \qquad (50)$$

As such, the PREs ascent iteration in Algorithm 1 remains unchanged, but the beamforming ascent iteration is processed differently, namely as follows.

Let as before $\tilde{f}_b^{(\kappa)}(\mathbf{W})$ be defined by (29). Define also[5]

$$\lambda^{(\kappa)} \triangleq \max_{k\in\mathcal{K}_E} \lambda_{\max}(\Psi_k^{(\kappa)}). \tag{51}$$

We now rewrite $\tilde{f}_b^{(\kappa)}(\mathbf{W})$ defined by (29) as

$$\begin{aligned}
\tilde{f}_b^{(\kappa)}(\mathbf{W}) &= a^{(\kappa)} + 2\sum_{k\in\mathcal{K}_E} \Re\{\langle \tilde{B}_k^{(\kappa)}, \mathbf{W}_k\rangle\} - \lambda^{(\kappa)} \sum_{k\in\mathcal{K}_E} ||\mathbf{W}_k||^2 \\
&\quad + \sum_{k\in\mathcal{K}_E} \langle (\mathbf{W}_k)^H \left(\lambda^{(\kappa)} I_{N_t} - \Psi_k^{(\kappa)}\right) \mathbf{W}_k\rangle \qquad (52)\\
&= a^{(\kappa)} + 2\sum_{k\in\mathcal{K}_E} \Re\{\langle \tilde{B}_k^{(\kappa)}, \mathbf{W}_k\rangle\} - \lambda^{(\kappa)} P \\
&\quad + \sum_{k\in\mathcal{K}_E} \langle (\mathbf{W}_k)^H \left(\lambda^{(\kappa)} I_{N_t} - \Psi_k^{(\kappa)}\right) \mathbf{W}_k\rangle \\
&\geq a^{(\kappa)} + 2\sum_{k\in\mathcal{K}_E} \Re\{\langle \tilde{B}_k^{(\kappa)}, \mathbf{W}_k\rangle\} - \lambda^{(\kappa)} P \\
&\quad + \sum_{k\in\mathcal{K}_E} \Big[ 2\Re\Big\{ \langle (W_k^{(\kappa)})^H \big(\lambda^{(\kappa)} I_{N_t} - \Psi_k^{(\kappa)}\big) \mathbf{W}_k\rangle \Big\} \\
&\quad - \langle (W_k^{(\kappa)})^H \big(\lambda^{(\kappa)} I_{N_t} - \Psi_k^{(\kappa)}\big) W_k^{(\kappa)}\rangle \Big] \\
&= \hat{a}^{(\kappa)} + 2\sum_{k\in\mathcal{K}_E} \Re\{\langle \hat{B}_k^{(\kappa)}, \mathbf{W}_k\rangle\} \qquad (53)\\
&\triangleq \hat{f}_b^{(\kappa)}(\mathbf{W}), \qquad (54)
\end{aligned}$$

where $\hat{a}^{(\kappa)} \triangleq a^{(\kappa)} - 2\lambda^{(\kappa)} P + \sum_{k\in\mathcal{K}_E} \langle (W_k^{(\kappa)})^H \Psi_k^{(\kappa)} W_k^{(\kappa)}\rangle$, and $\hat{B}_k^{(\kappa)} \triangleq \tilde{B}_k^{(\kappa)} + \left(\lambda^{(\kappa)} I_{N_t} - \Psi_k^{(\kappa)}\right) W_k^{(\kappa)}$, $k \in \mathcal{K}_E$. The function $\hat{f}_b^{(\kappa)}(\mathbf{W})$ is a tight minorant of $\tilde{f}_b^{(\kappa)}(\mathbf{W})$ over the nonconvex domain constrained by (49), because we have $\hat{f}_b^{(\kappa)}(W^{(\kappa)}) = \tilde{f}_b^{(\kappa)}(W^{(\kappa)})$. We thus seek an ascent point $W^{(\kappa+1)}$ satisfying (25) from solving the following problem of tight minorant maximization

$$\max_{\mathbf{W}} \ \hat{f}_b^{(\kappa)}(\mathbf{W}) \quad \text{s.t.} \quad (49), \tag{55}$$

which is not convex, but still admits the following closed-form solution

$$\begin{aligned}
W_k^{(\kappa+1)}(i,j) &= \frac{\sqrt{P/N_t}}{\mu^{(\kappa)}(i)} (\hat{B}_k^{(\kappa)}(i,j)), \\
&\quad j = 1,\ldots,N_r; i = 1,\ldots,N_t; k \in \mathcal{K}_E, \qquad (56)
\end{aligned}$$

with $\mu^{(\kappa)}(i) = \left(\sum_{k\in\mathcal{K}_E} \sum_{j=1}^{N_r} |\hat{B}_k^{(\kappa)}(i,j)|^2\right)^{1/2}$, $i = 1,\ldots,N_t$.

Next, we consider the following per-TA inequality constraint instead of (49):

$$\max_{\mathbf{W},\boldsymbol{\theta}} \ f(\mathbf{W},\boldsymbol{\theta}) \quad \text{s.t.} \quad (4), \tag{57a}$$

$$\sum_{k'\in\mathcal{K}_E} \sum_{j=1}^{N_r} |\mathbf{W}_{k'}(i)|^2 \leq P/N_t, i = 1,\ldots,N_t. \tag{57b}$$

With $\lambda^{(\kappa)}$ defined in (51), it follows from (52) that, we have:

$$\tilde{f}_b^{(\kappa)}(\mathbf{W}) \geq a^{(\kappa)} - \sum_{k\in\mathcal{K}_E} (W_k^{(\kappa)})^H \left(\lambda^{(\kappa)} I_{N_t} - \Psi_k^{(\kappa)}\right) W_k^{(\kappa)}$$

$$+ \tilde{\hat{f}}_b^{(\kappa)}(\mathbf{W}), \tag{58}$$

where

$$\begin{aligned}
\tilde{\hat{f}}_b^{(\kappa)}(\mathbf{W}) &\triangleq -\lambda^{(\kappa)} \sum_{k\in\mathcal{K}_E} ||\mathbf{W}_k||^2 + 2\sum_{k\in\mathcal{K}_E} \Re\{\langle \hat{B}_k^{(\kappa)}, \mathbf{W}_k\rangle\} \\
&= \sum_{i=1}^{N_t} \sum_{k\in\mathcal{K}_E} \sum_{j=1}^{N_r} \Big[ -\lambda^{(\kappa)} |\mathbf{W}_k(i,j)|^2 \\
&\quad + 2\Re\{(\hat{B}_k^{(\kappa)}(i,j))^* \mathbf{W}_k(i,j)\}\Big] \qquad (59)
\end{aligned}$$

and $\hat{B}_k^{(\kappa)}$ is defined in (53). The RHS of (58) provides a tight minorant for $\tilde{f}_b^{(\kappa)}(\mathbf{W})$ in the LHS as they match at $W^{(\kappa)}$. We thus seek an ascent point $W^{(\kappa+1)}$ satisfying (25) from solving the following problem of tight minorant maximization. We thus generate $W^{(\kappa+1)}$ by solving the problem

$$\max_{\mathbf{W}} \ \tilde{\hat{f}}_b^{(\kappa)}(\mathbf{W}) \quad \text{s.t.} \quad (57b), \tag{60}$$

which is decomposed into $N_t$ independent subproblems

$$\begin{aligned}
\max_{\mathbf{W}_k(i)} & \sum_{k\in\mathcal{K}_E} \sum_{j=1}^{N_r} \Big[ -\lambda^{(\kappa)} |\mathbf{W}_k(i,j)|^2 \\
&\qquad + 2\Re\{(\hat{B}_k^{(\kappa)}(i,j))^* \mathbf{W}_k(i,j)\}\Big] \\
\text{s.t.} \quad & \sum_{k\in\mathcal{K}_E} \sum_{j=1}^{N_r} |\mathbf{W}_k(i,j)|^2 \leq P/N_t, \qquad (61)
\end{aligned}$$

each of which admits the closed-form solution

$$W_k^{(\kappa+1)}(i,j) = \begin{cases} \dfrac{1}{\lambda^{(\kappa)}} \hat{B}_k^{(\kappa)}(i,j) & \text{if} \quad \bar{\Xi}^\kappa \leq (\lambda^{(\kappa)})^2 P/N_t \\ \dfrac{1}{\lambda^{(\kappa)} + \mu^{(\kappa)}(i)} \hat{B}_k^{(\kappa)}(i,j) & \text{otherwise,} \end{cases} \tag{62}$$

where $\bar{\Xi}^\kappa \triangleq \sum_{k\in\mathcal{K}_E} \sum_{j=1}^{N_r} |\hat{B}_k^{(\kappa)}(i,j)|^2$ and $\mu^{(\kappa)}(i)$ is found from bisection such that

$$\sum_{k\in\mathcal{K}_E} \sum_{j=1}^{N_r} |\hat{B}_k^{(\kappa)}(i,j)|^2 = (\lambda^{(\kappa)} + \mu^{(\kappa)}(i))^2 P/N_t.$$

Algorithm 2 recaps the development of this section. Like Algorithm 1, its convergence to a local solution satisfying the first order optimality condition is granted.

**Algorithm 2** Per-TA power constrained GM algorithm for (50)/(57).

1: **Initialization:** Set $\kappa = 0$. Generate $(w^{(0)}, \theta^{(0)})$ feasible for (50)/(57).
2: **Repeat until convergence of the objective in (50)/(57):** Generate $W^{(\kappa+1)}$ by (56)/(62), and $\theta^{(\kappa+1)}$ by (43). Reset $\kappa \leftarrow \kappa + 1$.
3: **Output** $(W^{(\kappa)}, \theta^{(\kappa)})$ and rates $r_k(W^{(\kappa)}, \theta^{(\kappa)})$, $k \in \mathcal{K}$ with their achieved GM $\Big[\prod_{k\in\mathcal{K}_1} \Big(r_k(W^{(\kappa)}, \theta^{(\kappa)}) \min_{i=1,2} \big\{r_{\pi(k)}(W^{(\kappa)}, \theta^{(\kappa)}) + r_{i,K+k}(W^{(\kappa)}, \theta^{(\kappa)})\big\}\Big)\Big]^{1/K}$.

[5]In fact, we can even avoid the eigenvalue calculation by taking any $\lambda^{(\kappa)}$ more than $\max_{k\in\mathcal{K}_E} \lambda_{\max}(\Psi_k^{(\kappa)})$ such as $\max_{k\in\mathcal{K}_E} \langle \Psi_k^{(\kappa)}\rangle/\mu$ for some $\mu > 1$.

## V. Numerical examples

In our simulations, we set the path-loss of the BS-to-UE $k$ link at a distance $d_{B,k}$ to $\beta_{B,k} \triangleq G_{\text{BS}} - 33.05 - 36.7\log_{10}(d_{B,k})$ (dB), that of the BS-to-RIS link at a distance $d_{B,R}$ to $\beta_{B,R} = G_{\text{RIS}} + G_{\text{BS}} - 35.9 - 22\log_{10}(d_{B,R})$ (dB), and that of the RIS-to-UE $k$ link at a distance $d_{R,k}$ to $\beta_{R,k} \triangleq G_{\text{RIS}} - 33.05 - 30\log_{10}(d_{R,k})$ (dB), where $G_{\text{RIS}} = 5$ dBi and $G_{\text{BS}} = 5$ dBi are the antenna gains of the BS and RIS elements. Similar to [5], [35], the above parameters are modeled using the 3GPP urban micro (UMi) scenario from [44, Table B.1.2.1-1] under 2.5 GHz operating frequency. The elements of the LoS channel matrix between the BS and RIS are given by $[H_{B,R}]_{n,m} = e^{j\pi\left((n-1)\sin\bar{\theta}_n\sin\bar{\phi}_n+(m-1)\sin\theta_n\sin\phi_n\right)}$, where $\theta_n$ and $\phi_n$ are uniformly distributed as $\theta_n \sim \mathcal{U}(0,\pi)$ and $\phi_n \sim \mathcal{U}(0,2\pi)$, respectively, and $\bar{\theta}_n = \pi - \theta_n$ and $\bar{\phi}_n = \pi + \phi_n$ [5]. The elements of normalized small-scale fading channel matrix $H_{B,k}$ follow Rayleigh distribution, while those of the small-scale fading channel matrix $H_{R,k}$ obey Rician distribution with a Rician K-factor of 3. The spatial correlation matrix, which models the correlation between RIS elements, is given by $[R_{R,k}]_{n,n'} = e^{j\pi(n-n')\sin\tilde{\phi}_k\sin\tilde{\theta}_k}$, where $\tilde{\phi}_k$ and $\tilde{\theta}_k$ represent the azimuth and elevation angle for UE $k$, respectively. The azimuth angles are generated using the Von Mises distribution with mean angle of departure (AoD) of $\bar{\tilde{\phi}}_k$ and spread of $0.2$, while the elevation angles are generated using the Laplace distribution with mean AoD $\bar{\tilde{\theta}}_k$ and spread of $8^o$ [37]. We set the noise power spectral density to $= -174$ dBm/Hz. Additionally, $N_t = 7$ antennas are employed at the BS and $K = 10$ UEs are assumed at random locations. Unless stated otherwise, we set $N = 100$ PREs at RIS, $b = 3$ bits for the quantized PREs, and $N_r = 2$ antennas at the UEs.

We use the following legends to specify the proposed implementations: "n-NOMA"/"NOMA"/ "CoSig" refers to the specific scheme, "$b = \infty/3$" refers to the resolution of RIS PREs, and "TA-equal-P"/"TA-ineq.-P" refers to the TA-wise equal power constraints (49)/TA inequality power constraints (57b), which are addressed by Algorithm 2. The default when the latter is absent is the sum power constraint, which is addressed by Algorithm 1. Lastly "ST max." refers to ST-maximization.

We consider a pair of practical scenarios for simulations. The first one assumes the availability of a direct links between the BS and UEs, as shown in Fig. 1, while the second scenario assumes that the direct transmission path between the BS and UEs is blocked by some obstruction [28]. We term them as "Scenario 1" and "Scenario 2", respectively. In the following subsections, we characterize the performance of the proposed algorithms under both scenarios.

### A. Results for Scenario 1

Under this scenario, which is shown in Fig. 1, the coordinate-locations of the BS and RIS are given by $(60, 0, 25)$ m and $(0, 90, 40)$ m, respectively, where $25$ m and $40$ m refers to the height of BS antennas and PREs of RIS from the ground, respectively. The UEs are randomly placed in a $180m \times 180m$ area to the right of the BS and RIS. Unless stated otherwise, the transmit power budget is set to $P = 16$ dBm.

Fig. 2 plots the achievable GM-throughput versus the transmit power budget $P$ for both $b = 3$-bit resolution and $\infty$-bit resolution PREs achieved by Algorithm 1. Fig. 2 shows a very small gap between their performances. This is due to the presence of a direct path between the BS and UEs. Fig. 2 clearly shows the performance gain of n-NOMA over NOMA and CoSig, which increases upon increasing $P$. Fig. 3 plots the achievable GM-throughput versus the number of receiver antennas $N_r$. Fig. 3 clearly shows the performance gain of n-NOMA and NOMA over CoSig, which increases upon increasing $N_r$. More particularly, in contrast to n-NOMA and NOMA, CoSig fails to provide any significant performance gain with the increase in $N_r$.

Fig. 4 plots the achievable ST of GM-throughput maximization (Algorithm 1) and compares it with the respective ST maximization (Section III-D). As expected, the achievable ST of the proposed GM-throughput maximization is lower than that of sum-throughput maximization. However, it can be observed from Fig. 4 that the achievable ST of the n-NOMA scheme is only modestly compromised (just $3\%$ drop in the ST at $P = 24$ dBm power budget). Note that a clear advantage of GM-throughput maximization over ST maximization is that the former avoids assigning a zero or low throughput to any UE and thus ensures low deviation among the users' throughput, which will be shortly shown through simulations. Actually, there is no need to enforce additional computationally intractable QoS throughput constraints in GM-throughput optimization due to the specific nature of its objective function (see e.g., (17)). On the other hand, the QoS throughput constraints have to be included in the ST maximization problem to avoid the assignment of a zero or low throughput to any UE. This QoS-constrained ST maximization problem has to be solved using a convex-solver based approach, which is computationally very complex compared to the proposed GM-throughput optimization. The detailed computational complexity of the proposed Algorithms and convex-solver based approach is provided in Section V-C.

Fig. 5 shows the effect of imperfect CSI on the achievable GM-throughput. In order to simulate this effect, we introduced random channel estimation errors both into the BS-to-UE channels and RIS-to-UE channels.[6] The magnitude of those channel estimation errors is bounded by $\delta$ times the magnitude of the corresponding channel estimates, where $\delta$ represents the relative CSI uncertainty [45]. As expected, we can observe from Fig. 5 that the GM-throughput decreases upon increasing the channel uncertainty. However, the reduction is quite modest. Explicitly, even at a very high channel uncertainty of $\delta = 0.08$, we observe a moderate $25\%$ drop in the GM-throughput of the n-NOMA scheme, which shows the robustness of our proposed algorithms against CSI uncertainty.

[6]We consider channel uncertainty only in the BS-to-UEs channels and RIS-to-UEs channels [45]. This is because it is convenient to obtain BS-to-RIS channel with a very small error compared to the BS-to-UEs and RIS-to-UEs channels [45].

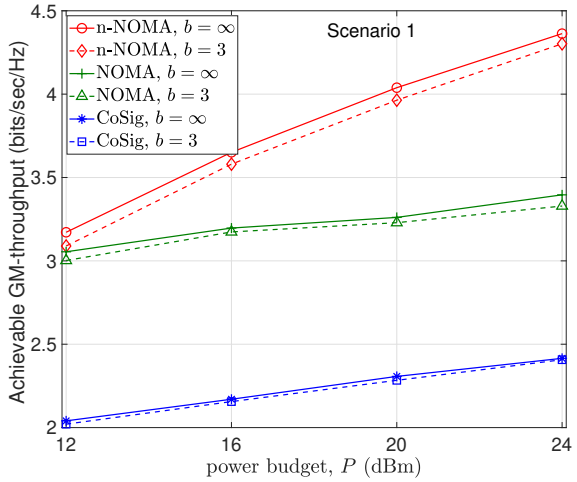


Fig. 2: Achievable users' GM-throughput versus the transmit power budget $P$ under Scenario 1.

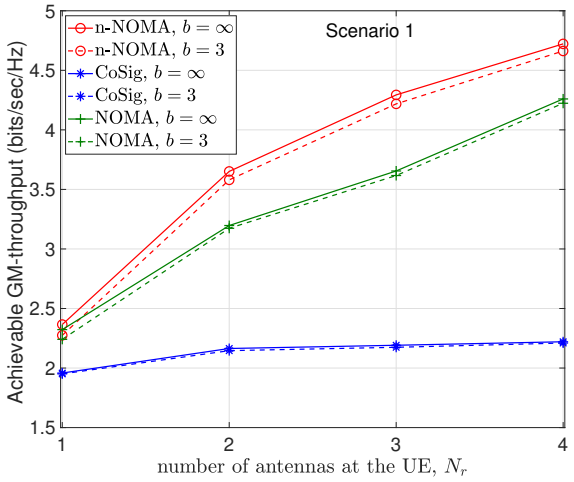


Fig. 3: Achievable users' GM-throughput versus number of receiver antennas $N_r$ under Scenario 1.

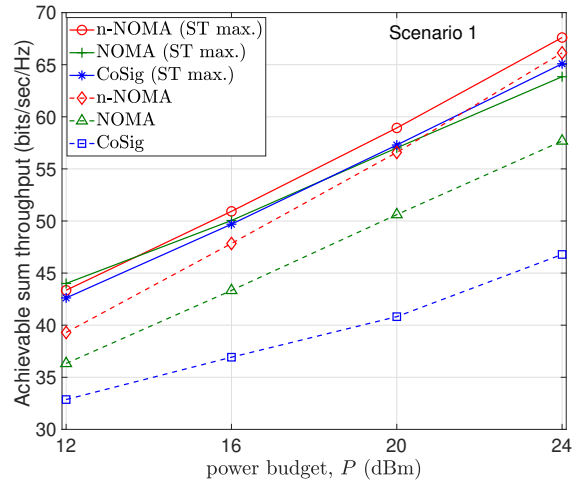


Fig. 4: Achievable sum-throughput versus the transmit power budget $P$ under Scenario 1.

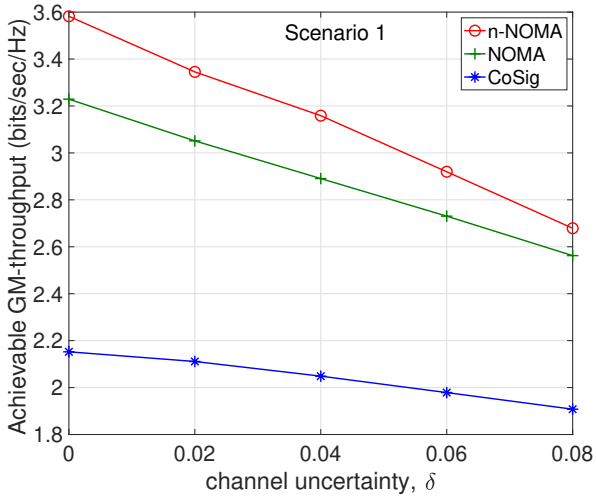


Fig. 5: Achievable users' GM-throughput versus relative CSI uncertainty $\delta$ under Scenario 1.

Fig. 6 plots the ratio between the maximum and minimum UE-throughput, $\frac{\max_k r_k}{\min_k r_k}$ versus $N_r$. Fig. 6 shows that the throughput-ratio is lower for NOMA and n-NOMA compared to CoSig, which shows the advantage of employing NOMA and n-NOMA in terms of a fairer user-throughput distribution. We can also observe from Fig. 6 that the throughput-ratio decreases by increasing $N_r$, which shows the benefit of deploying multiple antennas at the UE in achieving fairer user-throughput distribution.. Fig. 6 also shows that by employing ST maximization, the ratio between the maximum and minimum UE-throughput tends to $\infty$ because ST maximization may end up assigning zero or a very low throughput to some UEs. The advantage of deploying multiple antennas at the UE and employing n-NOMA and NOMA over CoSig in terms of promising a fairer user-throughput distribution can also be seen from Fig. 7, which shows the standard-deviation among the users' throughputs for these schemes.[7] The standard deviation among the users' throughput decreases upon increasing $N_r$. This is because when we increase $N_r$, the GM-throughput increases due to the increase in the number of receive antennas. Next, it can be observed from the definition of GM-throughput in (17) that maximization of GM-throughput actually reflects the improvement in the throughput of all the users, which reduces the standard deviation among the users' throughputs.

Fig. 8 plots Jain's fairness index (JFI) of the users' throughput, which is defined as JFI $= \frac{\left(\sum_{k=1}^{K} r_k\right)^2}{K \sum_{k=1}^{K} r_k^2}$ [46]. JFI is a well-known metric of quantifying the QoS fairness and it assumes continuous values in $[0, 1]$, where the value 1 implies maximum fairness (minimum standard-deviation) among the users' throughput [46]. It can be observed from Fig. 8 that JFI increases upon increasing $N_r$ and that the n-NOMA scheme achieves the highest JFI, outperforming both the NOMA and CoSig schemes. This result is equivalent to saying that the standard-deviation among the users' throughput decreases upon increasing $N_r$ and the n-NOMA scheme has a lower standard-deviation than the NOMA and CoSig schemes, which

[7]For fairness, the standard-deviation results in this paper are normalized by the mean of the users' throughput.

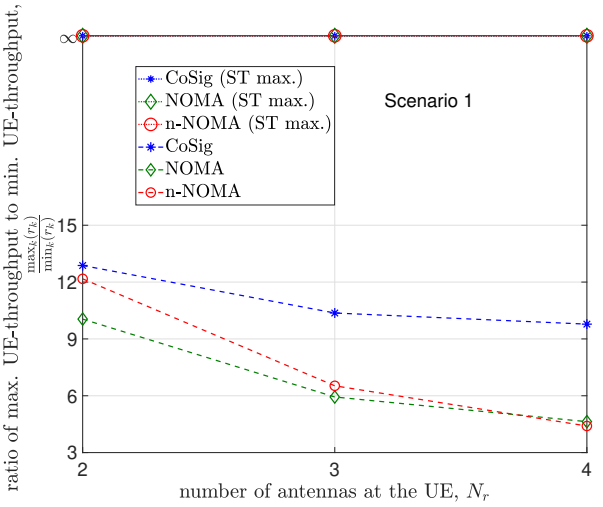


Fig. 6: Ratio between maximum and mini- mum UE-throughput versus $N_r$.

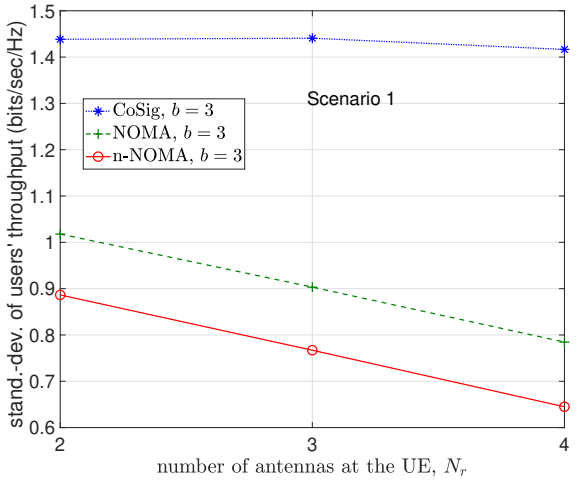


Fig. 7: Standard deviation among the users' throughputs versus $N_r$ under Scenario 1.

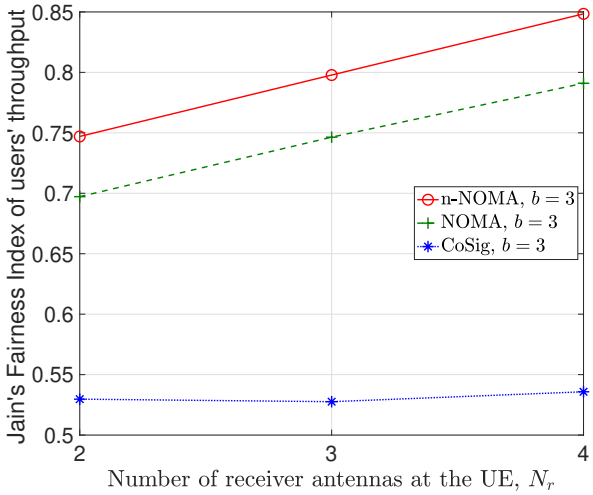


Fig. 8: Jain's fairness index of the users' throughputs versus $N_r$ under Scenario 1.

is consistent with what we observed in Fig. 7. In other words, these results prove that the n-NOMA attains the best throughput fairness outperforming both the NOMA and CoSig schemes and that the fairness increases with the increase of $N_r$.

Fig. 9 plots the achievable GM-throughput versus the power budget $P$ for both $b = 3$-bit resolution and $\infty$-bit resolution PREs under the per-TA equality power (TA-equal-P) constraints (49), which is computed by Algorithm 2. Again, the performance gap between the $b = 3$ and $b = \infty$ cases is small, due to the presence of a direct path between the BS and UEs under Scenario 1. Similar trend in the results is observed under per-TA inequality power constraints (57b). The performance comparison between the achievable GM-throughput under per-TA equality power constraints (49) and that under per-TA inequality power constraints (57b) versus the number of receiver antennas $N_r$ is provided in Fig. 10, which shows that the performances of (50) and (57) are similar, particularly for $N_r \geq 2$. This shows that the achievable GM-throughput is maximized if the TAs transmit at their maximum available power budget. Fig. 10 also shows the advantage of employing n-NOMA and NOMA over CoSig.

### *B. Results for Scenario 2*

In this subsection, we consider another practical situation, namely "Scenario-2", where the direct path between the BS and UEs is blocked by obstacles (i.e., $H_{B,k} \equiv 0$), as shown in Fig. 11. Due to the absence of the direct path, the distances between the nodes have to be kept smaller, i.e., the coordinate-location of he BS and RIS is given by $(20, 0, 25)$ m and $(0, 30, 40)$ m, respectively, while the UEs are randomly placed in a $85m \times 85m$ area to the right of the BS and RIS. Since the direct path between the BS and UEs is blocked, the UEs are paired under this scenario based on their geographical distance from the RIS. Unless stated otherwise, the transmit power budget is set to $P = 34$ dBm.

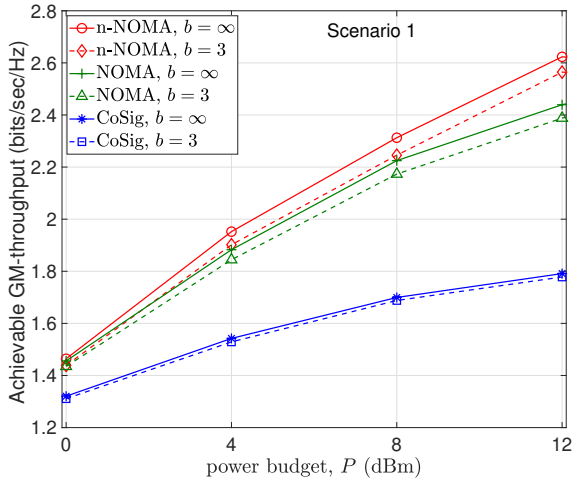


Fig. 9: Achievable users' GM-throughput under TA-wise equal-power constraint (49) versus $P$ under Scenario 1.

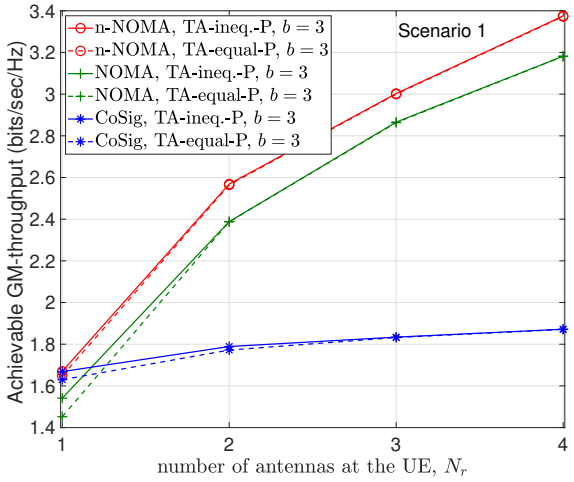


Fig. 10: Achievable users' GM-throughput under the TA-wise constraints (49) and (57b) under Scenario 1.

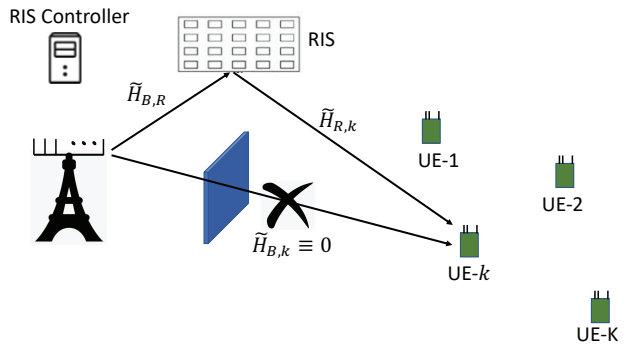


Fig. 11: System model under Scenario 2

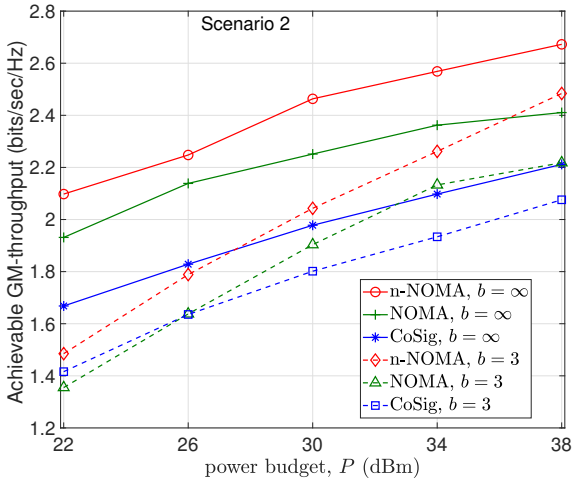


Fig. 12: Achievable users' GM-throughput versus the transmit power budget $P$ under Scenario 2.

Fig. 12 plots the achievable GM-throughput versus the transmit power budget $P$ for both $b = 3$-bit and $\infty$-bit resolution PREs. In contrast to Fig. 2 of Scenario 1, Fig. 12 clearly shows the gap between the performances of the $b = 3$-bit and $b = \infty$-bit resolutions, which is due to the absence of a direct path between the BS and UEs, and it decreases upon increasing $P$. Fig. 12 also shows the superiority of n-NOMA over NOMA and CoSig. The same trend is also observed in Figures 13, 14, and 15, which plot the achievable GM-throughput versus the number of receiver antennas $N_r$, the resolution of PREs $b$, and the number of PREs $N$.

"It can be observed from Figures 13, 14, and 15 that there is a wider gap between the performances of low- and high-resolution PREs in Scenario 2 than in Scenario 1. This is because in Scenario 1, the presence of a direct BS-UE link is the major source of throughput enhancement, which can be observed from the simulation results since the performance gap between the low- and high-resolution PREs is small. On the other hand, in Scenario 2, there is no direct link between the BS and UEs. The users' throughput is only dependent on the indirect twin-hop BS-RIS-UE link. In other words, the achievable throughput depends on the presence or absence of RIS. Hence, the resolution of RIS PREs has a substantial impact on the overall throughput and we see a clear gap between the performances of low- and high-resolution PREs."

Fig. 16 plots the achievable ST versus the number of PREs $N$ under Scenario 2 and compares it to that achieved by the ST maximization approach. As expected, the achievable ST of the proposed GM-throughput maximization (Algorithm 1) is lower than that by the ST maximization (Section III-D). The advantage of GM-throughput maximization is that it avoids assigning a zero or a very low throughput to any UE and thus ensures a smaller deviation among the users' throughputs.

Fig. 17 plots the standard deviation among the users' throughputs versus the number of PREs $N$. By reducing the standard deviation among the users' throughputs, Fig. 17 shows the advantage of employing n-NOMA and NOMA over CoSig in terms of promising a fairer user-throughput distribution. Figures 16 and 17 also show that by employing the pro-

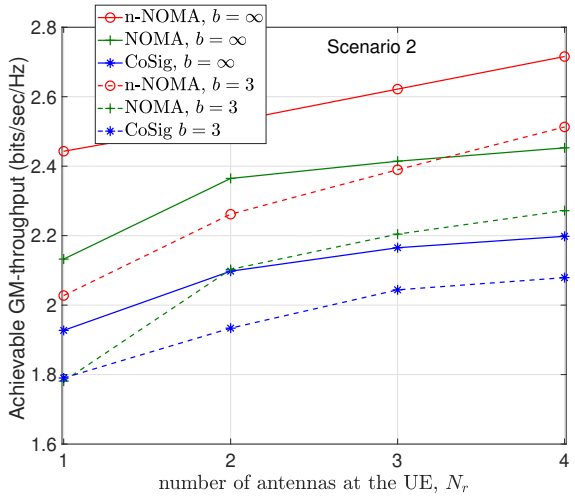


Fig. 13: Achievable users' GM-throughput versus number of receive-antennas $N_r$ under Scenario 2.

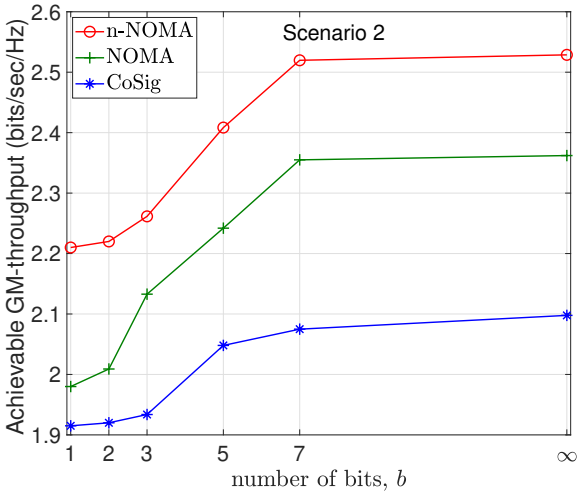


Fig. 14: Achievable users' GM-throughput versus the bit-resolution $b$ of PREs under Scenario 2.

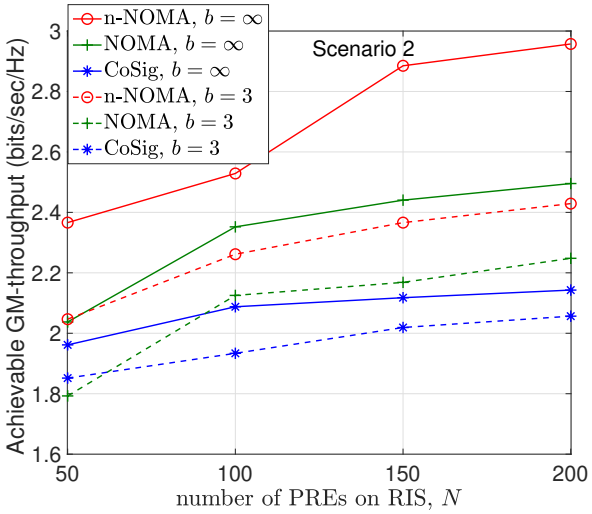


Fig. 15: Achievable users' GM-throughput versus number of PREs $N$ under Scenario 2.

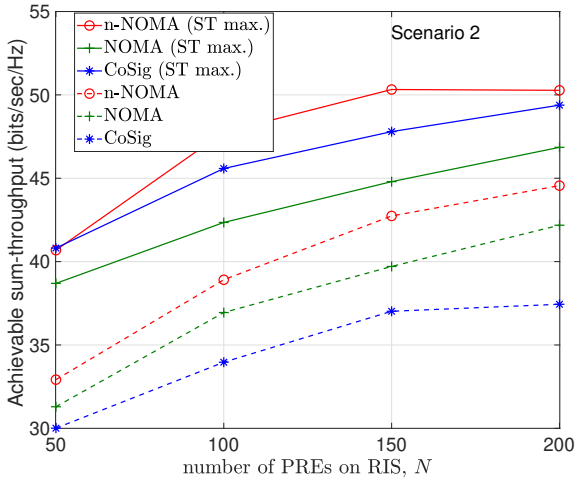


Fig. 16: Achievable sum-throughput versus number of PREs $N$ under Scenario 2.

posed GM-throughput maximization, increasing the number of PREs not only increases the achievable sum-throughput, but also slightly decreases the standard deviation among the users' throughputs. This is because when we increase the number of PREs, the GM-throughput increases due to the increase in diversity gain through additional reflected paths. Furthermore, it can be observed from the definition of GM-throughput in (17) that the maximization of the GM-throughput actually reflects the improvement in the throughput of all the users, which yields a reduced standard deviation among the users' throughputs.

Fig. 18 shows the effect of imperfect CSI on the achievable GM-throughput under Scenario 2. Observe from Fig. 18 that the GM-throughput decreases with the increase of the channel uncertainty $\delta$. However, in contrast to Scenario 1, which exhibits a moderate drop of $25\%$ in the throughput, the GM-throughput drops only by $4\%$ under Scenario 2 at a high channel uncertainty of $\delta = 0.08$. This is because the BS-to-UEs links do not exist in Scenario 2 and the channel uncertainty only contaminates the RIS-to-UEs links.

Figures 19 and 20 plot the achievable GM-throughput and the standard deviation among the users' throughput, under both the per-TA equality power constraints (49) and inequality power constraints (57b). We observe that (57) achieves better performance than (50) because the former does not limit each BS TA to transmit at the fixed power $P/N_t$. In contrast to all the results observed under Scenario 1 and that observed under Scenario 2 under the sum-power constraint, these figures show the supremacy of CoSig and n-NOMA over NOMA. This may be explained by realizing that in Scenario 2, no diversity contribution is gleaned from the direct BS-UE link. Under this situation, if we enforce the per-TA power constraints, the n-NOMA scheme does not get much room for intelligent resource allocation.

### C. Computational Complexity

The proposed algorithms are computationally efficient, since their solution is based on evaluating closed-form expressions

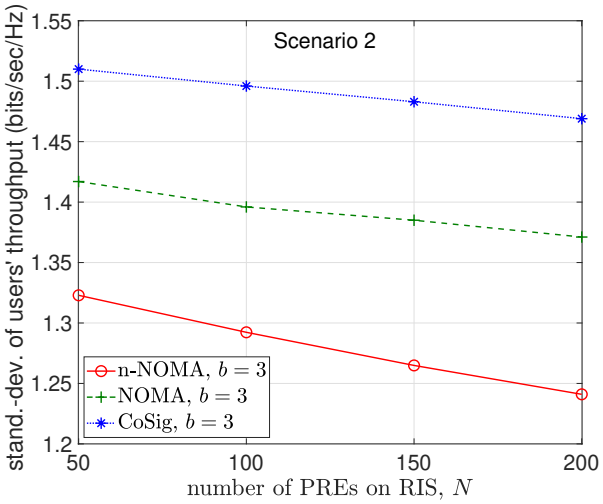


Fig. 17: Standard deviation among the users' throughputs versus the number of PREs $N$ under Scenario 2.

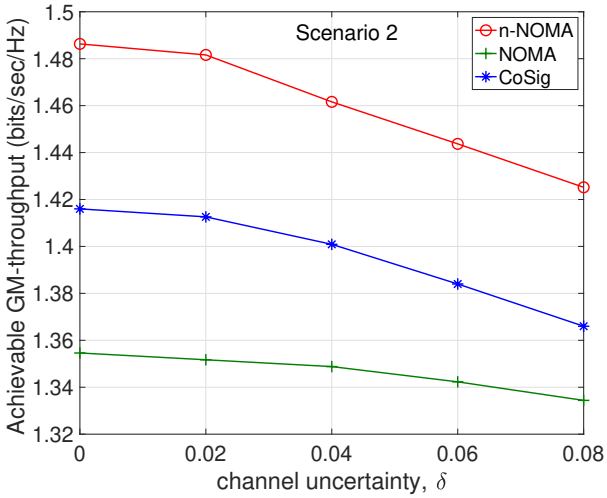


Fig. 18: Achievable users' GM-throughput versus relative CSI uncertainty $\delta$ under Scenario 2.

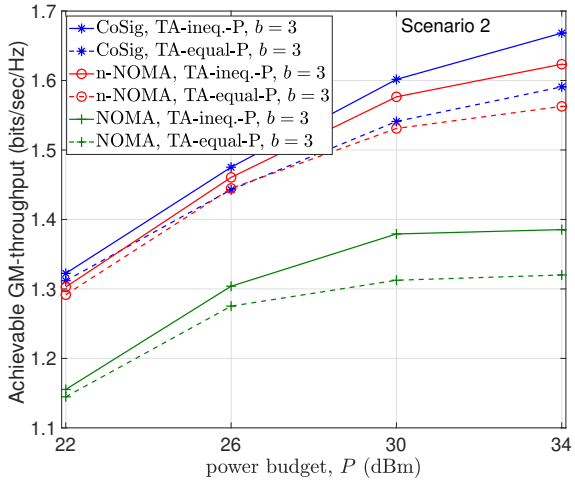


Fig. 19: Achievable users' GM-throughput under the TA-wise constraints (49) and (57b) under Scenario 2.

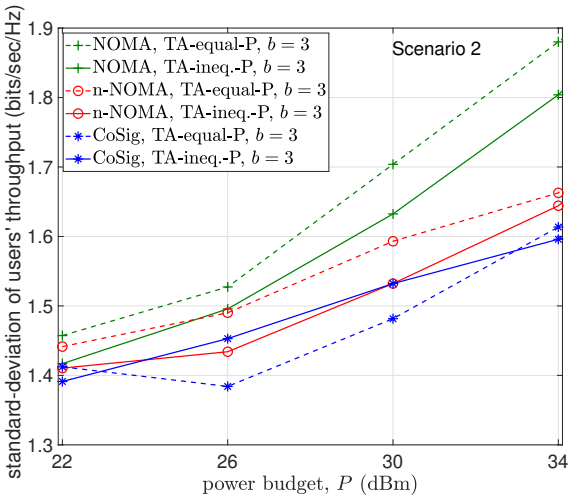


Fig. 20: Standard deviation among the users' throughputs under the TA-wise constraints (49) and (57b).

TABLE II: Average number of iterations required for the convergence of proposed algorithms.

| | Scenario 1 | | | Scenario 2 | | |
|---|---|---|---|---|---|---|
| Algorithms | Alg. 1 | Alg. 2 for (50) | Alg. 2 for (57) | Alg. 1 | Alg. 2 for (50) | Alg. 2 for (57) |
| CoSig | 24.45 | 42.96 | 41.93 | 16.6 | 36.4 | 44.15 |
| NOMA | 33.95 | 94.13 | 93.06 | 31.36 | 74.8 | 96.1 |
| n-NOMA | 44.75 | 91.38 | 91.12 | 27.74 | 45.75 | 54.65 |

at each iteration. The number of iterations required for the convergence of the proposed algorithms is shown in Table II. It can be observed from Table II that generally, n-NOMA requires more iterations to converge than CoSig and NOMA, because the former has to optimize one and a half times higher number of beamforming matrices compared to the latter. Table II also shows that Algorithms 2 requires more iterations for convergence compared to Algorithm 1. This is because the former involves $M$ power constraints (at each TA) instead of a single sum-power constraint.

Table III shows the computational complexity of the proposed Algorithms 1-3 and compares it to that of a convex-solver based approach for the considered three problems (i) sum-power (sum-P) constraint (const.) based optimization (opt.) in (19), (ii) per-TA equality power (equality-P) const. based opt. in (50), and (iii) per-TA inequality power constraint based optimization in (57). It can be observed from Table III that the proposed Algorithms 1-3 are computationally more efficient than the convex-solver based approach, because our proposed solution is based on closed-form expressions. Moreover, we would like to mention that a convex-solver based approach to jointly designing the BS's transmit beamformers and RIS PREs in our RIS-enabled MU MIMO-NOMA system is not available in the open literature, so fair performance

TABLE III: Computational complexity of the proposed Algorithms and convex-solver based approach.

| Problems | Proposed Algorithms | | Convex-solver based approach [10] | |
|---|---|---|---|---|
| | BF iteration | PREs iteration | BF iteratiion | PREs iteration |
| sum-P const. based opt. (19) | $\mathcal{O}\left(N_t \log_2(N_t) N_r K\right)$ | $\mathcal{O}(N)$ | $\mathcal{O}\left((N_t N_r K)^3\right)$ | $\mathcal{O}(N^3)$ |
| per-TA equality-P const. based opt. (50) | $\mathcal{O}\left(N_t N_r K\right)$ | $\mathcal{O}(N)$ | $\mathcal{O}\left((N_t N_r K)^3 N_t\right)$ | $\mathcal{O}(N^3)$ |
| per-TA inequality-P const. based opt. (57) | $\mathcal{O}\left(N_t N_r K\right)$ | $\mathcal{O}(N)$ | $\mathcal{O}\left((N_t N_r K)^3 N_t\right)$ | $\mathcal{O}(N^3)$ |

comparisons to the existing solutions cannot be provided.

## VI. Conclusions

This paper has considered the joint design of transmit beamformers at the BS and PREs at the RIS for RIS-aided MU MIMO networks to improve the throughput fairness of all users, under both the sum-power and per-TA power constraints. A new-NOMA (n-NOMA)-based signaling scheme has been adopted, which includes both NOMA and CoSig as particular cases. The proposed solution is based on GM-throughput maximization, which iterates by evaluating closed-form expressions of very low computational complexity. Thus, the proposed design is eminently suitable for very large networks, and GM-throughput maximization improves all users' throughputs, resolving the issue of an unfair throughput allocation inherent in sum-throughput maximization. Our extensive simulations results have shown the performance advantage of n-NOMA and NOMA-based RIS implementation over CoSig in terms of improving the achievable GM-throughput and minimizing the standard deviation among the users' throughputs. However, for a particular scenario, where direct communication between the BS and UEs is blocked by obstacles, it has been shown through simulations that NOMA is outperformed by CoSig under per-TA power constraints based design.

## References


[1] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wirel. Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[2] M. Di Renzo *et al*., "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and road ahead," *IEEE J. Sel. Areas Commun*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[3] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. D. Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surv. & Tut.*, vol. 23, no. 3, pp. 1546–1577, thirdquarter 2021.

[4] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[5] Q. U. A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M. S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.

[6] G. Zhou, C. Pan, H. Ren, K. Wang, W. Xu, and A. Nallanathan, "Intelligent reflecting surface aided multigroup multicast MISO communication systems," *IEEE Trans. Signal Process.*, vol. 68, pp. 3236–3251, 2020.

[7] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surface," *IEEE Trans Wirel. Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

[8] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[9] M. M. Zhao, Q. Wu, M. J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 1, pp. 2–17, Jan. 2021.

[10] H. Yu, H. D. Tuan, A. A. Nasir, T. Q. Duong, and H. V. Poor, "Joint design of reconfigurable intelligent surfaces and transmit beamforming under proper and improper Gaussian signaling," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2589–2603, Nov. 2020.

[11] H. Yu, H. D. T. E. Dutkiewicz, H. V. Poor, and L. Hanzo, "Maximizing the geometric mean of user-rates to improve rate-fairness: Proper vs. improper Gaussian signaling," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 1, pp. 295–309, Jan. 2022.

[12] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Techn. Conf. (VTC Spring)*, 2013, pp. 1–5.

[13] Z. Ding et al., "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[14] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, thirdquarter 2018.

[15] V. D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and O. S. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Select. Areas Commun.*, vol. 35, no. 12, pp. 2681–2695, Dec. 2017.

[16] H. D. Tuan, A. A. Nasir, and M.-N. Nguyen, "Han-Kobayshi signaling in MIMO broadcasting," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 855–858, May 2019.

[17] A. A. Nasir, H. D. Tuan, H. H. Nguyen, T. Q. Duong, and H. V. Poor, "Signal superposition in NOMA with proper and improper Gaussian signaling," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6537–6551, Oct. 2020.

[18] Z. Ding and H. V. Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, May 2020.

[19] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "Reconfigurable intelligent surface aided NOMA networks," *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2575–2588, Nov. 2020.

[20] Y. Cheng, K. H. Li, Y. Liu, K. C. Teh, and G. K. Karagiannidis, "Non-orthogonal multiple access (NOMA) with multiple intelligent reflecting surfaces," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 11, pp. 7184–7195, Nov. 2021.

[21] Y. Cheng, K. H. Li, Y. Liu, K. C. Teh, and H. V. Poor, "Downlink and uplink intelligent reflecting surface aided networks: NOMA and OMA," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 6, pp. 3988–4000, June 2021.

[22] Z. Tang, T. Hou, Y. Liu, J. Zhang, and C. Zhong, "A novel design of RIS for enhancing the physical layer security for RIS-aided NOMA networks," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 11, pp. 2398–2401, Nov. 2021.

[23] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Capacity and optimal resource allocation for IRS-assisted multi-user communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3771–3786, Jun. 2021.

[24] J. Zuo, Y. Liu, Z. Qin, and N. Al-Dhahir, "Resource allocation in intelligent reflecting surface assisted NOMA systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7170–7183, Nov. 2020.

[25] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 10, pp. 6884–6898, Oct. 2020.

[26] M. Fu, Y. Zhou, Y. Shi, and K. B. Letaief, "Reconfigurable intelligent surface empowered downlink non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3802–3817, Jun. 2021.

[27] P. Liu, Y. Li, W. Cheng, X. Gao, and X. Huang, "Intelligent reflecting surface aided NOMA for millimeter-wave massive MIMO with lens antenna array," *IEEE Trans. Vehic. Techn.*, vol. 70, no. 5, pp. 4419–4434, May 2021.

[28] Z. Zhang, C. Zhang, C. Jiang, F. Jia, J. Ge, and F. Gong, "Improving physical layer security for reconfigurable intelligent surface aided NOMA 6G networks," *IEEE Trans. Vehic. Techn.*, vol. 70, no. 5, pp. 4451–4463, May 2021.

[29] A. H. Phan, H. D. Tuan, H. H. Kha, and D. T. Ngo, "Nonsmooth optimization for efficient beamforming in cognitive radio multicast transmission," *IEEE Trans. Signal Process.*, vol. 60, pp. 2941–2951, June 2012.

[30] A. A. Nasir, H. D. Tuan, T. Q. Duong, and H. V. Poor, "UAV-enabled communication using NOMA," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5126–5138, July 2019.

[31] H. H. M. Tam, H. D. Tuan, and D. T. Ngo, "Successive convex quadratic programming for quality-of-service management in full-duplex MU-MIMO multicell networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2340–2353, June 2016.

[32] H. Tuy, *Convex Analysis and Global Optimization (second edition)*. Springer International, 2017.

[33] M. Di Renzo, M. Debbah, and et al., "Smart radio environments empowered by AI reconfigurable meta-surfaces: An idea whose time has come," *EURASIP J. Wirel. Commun. Network.*, no. 1, p. 129, May 2019.

[34] O. Ozdogan, E. Bjornson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *https://arxiv.org/abs/1911.03359*, vol. 9, no. 5, pp. 581–585, May 2020.

[35] E. Bjornson, O. Ozdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wirel. Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.

[36] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Tran. Commun.*, vol. 69, no. 6, pp. 4144–4157, June 2021.

[37] Q.-U.-A. Nadeem, A. Kammoun, M. Debbah, and M.-S. Alouini, "A generalized spatial correlation model for 3D MIMO channels based on the Fourier coefficients of power spectrums," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3671–3686, Jul. 2015.

[38] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[39] H. Dahrouj and W. Yu, "Multicell interference mitigation with joint beamforming and common message decoding," *IEEE Trans. Commun.*, vol. 59, no. 8, pp. 2264 –2273, Aug. 2011.

[40] E. Che, H. D. Tuan, H. H. M. Tam, and H. H. Nguyen, "Successive interference mitigation in multiuser MIMO interference channels," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2185–2199, June 2015.

[41] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 98–105, May 2016.

[42] H. Tuy and H. D. Tuan, "Generalized S-lemma and strong duality in nonconvex quadratic programming," *J. Global Opt.*, vol. 56, pp. 1045–1072, July 2013.

[43] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, Aug. 1978.

[44] 3GPP TR 36.814 V9.0.0, "Further advancements for evolved universal terrestrial radio access (E-UTRA) physical layer aspects," Mar. 2010.

[45] G. Zhou, C. Pan, H. Ren, K. Wang, M. D. Renzo, and A. Nallanathan, "Robust beamforming design for intelligent reflecting surface aided MISO communication systems," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 10, pp. 1658–1662, Oct. 2020.

[46] R. Jain, D.-M. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *Digital Equipment, Tech. Rep. DEC-TR-301*, Sept. 1984.